BOOTSTRAPPING IN A HIGH DIMENSIONAL BUT VERY LOW SAMPLE SIZE PROBLEM

A Dissertation

by

JUHEE SONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2005

Major Subject: Statistics

BOOTSTRAPPING IN A HIGH DIMENSIONAL BUT VERY LOW SAMPLE SIZE PROBLEM

A Dissertation

by

JUHEE SONG

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Jeffrey D. Hart (Chair of Committee)	
F. Michael Speed (Member)	Naisyin Wang (Member)
Joel Zinn (Member)	Michael T. Longnecker (Head of Department)

May 2005

Major Subject: Statistics

ABSTRACT

Bootstrapping in a High Dimensional
But Very Low Sample Size Problem. (May 2005)
Juhee Song, B.S., Inha University;
M.A., Yonsei University
Chair of Advisory Committee: Dr. Jeffrey D. Hart

High Dimension, Low Sample Size (HDLSS) problems have received much attention recently in many areas of science. Analysis of microarray experiments is one such area. Numerous studies are on-going to investigate the behavior of genes by measuring the abundance of mRNA (messenger RiboNucleic Acid), gene expression. HDLSS data investigated in this dissertation consist of a large number of data sets each of which has only a few observations.

We assume a statistical model in which measurements from the same subject have the same expected value and variance. All subjects have the same distribution up to location and scale. Information from all subjects is shared in estimating this common distribution.

Our interest is in testing the hypothesis that the mean of measurements from a given subject is 0. Commonly used tests of this hypothesis, the t-test, sign test and traditional bootstrapping, do not necessarily provide reliable results since there are only a few observations for each data set.

We motivate a mixture model having C clusters and $3C$ parameters to overcome the small sample size problem. Standardized data are pooled after assigning each data set to one of the mixture components. To get reasonable initial parameter estimates when density estimation methods are applied, we apply clustering methods

including agglomerative and K-means.

Bayes Information Criterion (BIC) and a new criterion, WMCV (Weighted Mean of within Cluster Variance estimates), are used to choose an optimal number of clusters.

Density estimation methods including a maximum likelihood unimodal density estimator and kernel density estimation are used to estimate the unknown density. Once the density is estimated, a bootstrapping algorithm that selects samples from the estimated density is used to approximate the distribution of test statistics. The t-statistic and an empirical likelihood ratio statistic are used, since their distributions are completely determined by the distribution common to all subject. A method to control the false discovery rate is used to perform simultaneous tests on all small data sets.

Simulated data sets and a set of cDNA (complimentary DeoxyriboNucleic Acid) microarray experiment data are analyzed by the proposed methods.

ACKNOWLEDGEMENTS

When I have thought of my graduate study in Texas A&M, I have admitted that I could not have finished it without the support of my family, my friends and professors in the Department of Statistics.

First, I would like to express profound gratitude to my advisor, Dr. Jeffrey D. Hart, for his patience, encouragement and insightful advising throughout this dissertation. I also express my appreciation to Dr. F. Michael Speed, Dr. Naisyin Wang and Dr. Joel Zinn for their kind suggestions and support as members of my advisory committee.

I am grateful to the faculty at Texas A&M who are willingly to support students. Many thanks to all the staff in the Department of Statistics, especially Ms. Marilyn Randall, for their kind service. I also thank my colleagues.

I thank my parents and parents-in-law for their support and everyday prayer to God for me. I also thank my sisters and brothers, Bohye, Minjung, Minhey, Jongbok, Yujin and Seokchul. I am most thankful for my family in Aggieland, my sincere husband, Jongil Lim, who is by my side, my precious son, Euntaek Daniel Lim, who always makes me smile and my unborn daughter who does not have a name yet. I accomplished this goal with their care and love. Finally, I thank God who loves me and provides me everything that I need.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

High dimensional, low sample size (HDLSS) data are produced in many areas of science, including chemometrics, microarray experiments and medical imaging (Hall et al., 2003). The data are composed of a large number of small data sets that have few observations. Hence, the dimension of the data vectors in HDLSS data is much larger than the size of each data set (Hall et al., 2003) since repetitions are too expensive to perform in related experiments.

Many statistical theories and methodology can be applied to such data to summarize them, make an inference or to test a hypothesis. One of the interesting problems with HDLSS data is testing a hypothesis on each of a large number of small data sets. For example, we may want to perform a test that the population mean of data set i is 0. Since data set i does not have many repetitions, usual tests such as t-tests, sign tests, permutation tests, and traditional bootstrapping do not necessarily provide valid results. Throughout this dissertation we explain methods to overcome the lack of repetitions in each data set.

One of the most active areas in HDLSS data is that of microarrays. Microarrays provide measures of the abundance of mRNA (messenger RiboNucleic Acid), which is the form of RNA that carries genetic information (Hall et al., 2003). Studies on the structure of the DNA (DeoxyriboNucleic Acid) of living creatures and the relationship between genes and certain diseases are very popular, and many useful properties

---

The format and style follow that of *Journal of the American Statistical Association*.

of genes are discovered from microarray experiments. The abundance of mRNA, or gene expression in a cell or gene, is related to the state of the cell or gene. Hence, gene expression is measured to obtain information about the activity of the cell or gene (Efron et al., 2001).

Microarray experiments generate numerics which correspond to the expression of each of the genes under various conditions. Statistical methodology can be applied to data from microarrays to infer common distributional properties of measurements (Efron et al., 2001) and to summarize the data.

We give a brief explanation of the steps of microarray experiments, since this dissertation analyzes data from a microarray experiment as an example of HDLSS data. There are two types of microarray experiments: On-chip oligonucleotide synthesis and spotted cDNA microarrays, the second of which we analyze in the dissertation. On-chip oligonucleotide synthesis deals with more genes than spotted cDNA microarrays.

The steps in a microarray experiment (Conzone and Pantanot, 2004) are as follows.

- Isolating target and labeling : Two separate RNA samples are extracted from an organism's tissue. One is referred to as reference and the other target. Reference and target might, for example, be from healthy and cancerous tissue, respectively. The two RNA samples are labeled with dissimilar fluorescent dyes.

- Hybridization : The labeled reference and target samples are combined and applied to the surface of a DNA microarray. For a given spot, if the target sample contains a cDNA which is complementary to the DNA in the spot, then the DNA will bind to the spot and the binding is called hybridization.

- Image scanning and quantifying: Noncomplementary targets and probes are

removed from the array surface by washing. Image scanning of the hybridized array is then conducted by a fluorescent reader or auto radiography to quantify the signal intensity.

- Database building, cleaning and normalization : Normalization is a procedure to remove the variation among slides so that one may compare gene expressions from different slides.

- Statistical analysis.

This dissertation is concerned with methodology that can be applied to the statistical analysis of microarray data. The methods dealt with include cluster analysis (Eisen et al., 1998; Fraley and Raftery, 2002; McLachlan et al., 2002), density estimation and bootstrapping (Van del Laan and Bryan, 2001). Various statistical methodologies including Bayesian models (Efron et al., 2001; Ibrahim et al., 2002; Ishwaran and Rao, 2003) can be used to analyze microarray data. The method of false discovery rates (FDRs) (Benjamini and Hochberg, 1995; Storey, 2003) is applicable since microarray experiments deal with simultaneous tests on many genes.

In the remainder of this chapter we discuss in more detail the type of HDLSS data to be considered and the problems addressed in this dissertation.

1.1 Descriptions of High Dimension, Low Sample Data

Our interest is in situations where a large number of subjects is available, but only a few readings are obtained from each subject, which is an example of HDLSS data. One can find this type of data in microarray experiments, which deal with many genes at the same time with few arrays.

A data set obtained from Dr. Kenneth Ramos and Dr. Charlie D. Johnson, both formerly of Texas A&M University, provides an example of such data. More than

2000 gene expressions of 813 genes under treatment and control conditions were measured. Only three expressions were measured for most genes.

The main interest in this problem is to determine which genes significantly express more under treatment than control. If we assume a model as in the next section, the problem of comparing gene expressions under treatment with gene expressions under control can be expressed as a hypothesis $H_0 : \beta_i = 0$, where $i = 1, \ldots, G$. Gene i has n_i repetitions, and G is much larger than any n_i.

We could use an ordinary t-test to test each H_0, but doing so could lead to an invalid test. If the observations are not normally distributed, then three or four observations are not enough to ensure approximate validity of the t-test. One could also use a nonparametric test, such as a sign test, that requires fewer distributional assumptions than the t-test. However, the power of such tests is not necessarily good with so few observations. One might consider bootstrapping in the traditional way by selecting bootstrap samples from the data for a single gene, but this does not seem feasible since there are only a few observations for each gene.

A new method is desirable to overcome the small sample size problem, and it is briefly introduced in the next section.

1.2 The Basic Problem

We observe L_{ij}, the jth measurement on the ith subject, that, for example, might be $L_{ij} = \log(T_{ij}/C_{ij})$, where T_{ij} and C_{ij} are treatment and control measurements, respectively. We assume that the following model holds:

$$L_{ij} = \beta_i + \gamma_i \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G,$$

where all ϵ_{ij} are i.i.d. from an unknown density f that has mean 0 and variance 1.

We wish to test the null hypotheses $H_{0i} : \beta_i = 0$ for $i = 1, \ldots, G$. The hypoth-

esis $\beta_i = 0$ means that measurements from the ith subject are not affected by the treatment.

In the previous section, we mentioned possible tests for the hypothesis $\beta_i = 0$, and noted the problem associated with each test. To mitigate the problem of few observations and to obtain reasonable tests, we use a mixture model consisting of several groups of subjects such that within a group subjects have similar means and variances. We may pool information from subjects in the same group in order to estimate f, the density of each ϵ_{ij}. It may be reasonable to assume that $\beta_i = 0$ for all subjects in one of the clusters, because we expect that relatively few subjects are affected by the treatment.

A bootstrap algorithm is adopted to approximate the sampling distribution of a test statistic. By using a statistic whose distribution is invariant to β_i and γ_i, the statistic's distribution is completely determined by the density f. To avoid identifiability problems we assume that f is unimodal.

Now our main concern is to get a reasonable estimate of f and/or the cumulative distribution function (CDF) F corresponding to f. We will assume that there are C distinct values of each of β_i and γ_i, where $C << G$. These will be denoted $\mu_1, \ldots, \mu_C$ and $\sigma_1, \ldots, \sigma_C$, respectively.

Estimation of f then proceeds in three steps.

(1) Apply a clustering algorithm with tentative value $\tilde{C}$ for C. The clustering algorithm provides estimates of means and standard deviations for each of the $\tilde{C}$ clusters, and of the conditional probability that any given subject belongs to any given cluster.

(2) Estimate the true number of clusters. We investigate two methods for doing so: BIC and a procedure that attempts to match the weighted average of within

cluster variance estimates ($WMCV$) with the average of G within subject variance estimates (MSV).

(3) Given the clustering corresponding to the estimate of C in (2), estimate f (or F) as described below.

We consider two density estimation methods: the unimodal density estimator (Wegman, 1969) and a modification of the ordinary kernel density estimator. A crucial problem with either of these methods is deciding which cluster a given subject belongs to.

For the unimodal density estimation scheme, we employ a clustering algorithm such as agglomerative or K-means (Hartigan, 1975; Tavazoie et al., 1999) to determine a definite cluster for each subject. Then standardized assignment data are defined as follows:

$$e_{ij} = \frac{L_{ij} - \hat{\mu}_{k(i)}}{\hat{\sigma}_{k(i)}}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G, \tag{1.1}$$

where $k(i)$ denotes the cluster to which subject i has been assigned and $\hat{\mu}_{k(i)}$ and $\hat{\sigma}_{k(i)}$ are the estimated mean and standard deviation, respectively, for that cluster. The unimodal density estimator (Wegman, 1969) is then computed from the data e_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, G$.

The kernel density estimation scheme differs from that just described in that each subject is assigned a probability of membership in each of the clusters. Defining

$$e_{ij(k)} = \frac{L_{ij} - \hat{\mu}_k}{\hat{\sigma}_k}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G,$$

the kernel estimator of f is

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{G} \sum_{k=1}^{C} \hat{\alpha}_{ki} \sum_{j=1}^{n_i} K\left(\frac{x - e_{ij(k)}}{h}\right), \tag{1.2}$$

where $N = \sum_{i=1}^{G} n_i$ and $\hat{\alpha}_{ki}$ is an estimate of the conditional probability, α_{ki}, that subject i belongs to cluster k given the observations $L_{i1}, \ldots, L_{in_i}$. The quantity α_{ki} depends upon f and $3C$ parameters: $\mu_1, \ldots, \mu_C$, $\sigma_1, \ldots, \sigma_C$ and $w_1, \ldots, w_C$, where

$$w_k = \text{proportion of subjects in population falling in cluster } k, \quad k = 1, \ldots, C.$$

Of course, f and the $3C$ parameters are unknown. To address this problem, an initial estimate of the matrix $[\alpha_{ki}]$ of conditional probabilities is obtained by one of three methods.

One method is to use a given f (such as a standard normal density) and estimate the $3C$ unknown parameters by fitting a mixture model via the EM algorithm. A second approach is to use an agglomerative clustering algorithm, to partition the subjects into clusters. This leads to simple estimates of $\mu_1, \ldots, \mu_C$ and $\sigma_1, \ldots, \sigma_C$. A third way is to use K-means clustering to estimate $\mu_1, \ldots, \mu_C$ and $\sigma_1, \ldots, \sigma_C$.

With these estimates, the data may be standardized as in (1.1), and then an ordinary estimate computed from all N standardized data values. Now, one may estimate the matrix $[\alpha_{ki}]$ and compute a kernel estimator of the form (1.2). This process could be iterated in more than one way to obtain other estimates of f, but we delay discussion of the details until chapter II.

Once we have an estimate of f, we test each hypothesis $H_0 : \beta_i = 0$ by using a bootstrap algorithm in which samples of size n_i are drawn from $\hat{f}$. Since we use a statistic whose sampling distribution only depends on n_i and f, we only have to perform this resampling for each distinct sample size among the subjects. In other words, the test statistics for two subjects with the same sample size may be compared to the same bootstrap percentile.

We consider two types of test statistics: the ordinary t-statistic and a likelihood ratio statistic using an estimate of f in place of f. The distribution of the t-statistic

and the likelihood ratio statistic are completely determined by f, the density of ϵ_{ij}, as shown in chapter III. Therefore getting an estimate of f is crucial to getting the distribution of the test statistic under the null hypothesis.

The key idea is that we borrow information from subjects which are in the same group, or cluster, in order to estimate f and to overcome the small sample size problem. The false discovery rate (FDR) method is used to deal with the problem of performing a multiplicity of tests simultaneously.

In the next chapter we will describe the estimation of f, which involves clustering methods and kernel density estimation. We also address the question of consistency of the estimated f.

1.3 Statistical Hypothesis Tests in Microarray Analyses

In this section we discuss some hypothesis testing methods that are commonly used on microarray data.

The simplest way to decide if a certain gene is affected by a treatment is the fold change method. The average of ratios of treatment and control measurements is compared with an arbitrary cut-off value to decide if a gene is significantly expressed or not (Cui and Churchill, 2003). This is not a formal statistical test since it does not involve any statistical distributions.

The ordinary t-test is also commonly used in microarray studies. The global t-statistic that uses the global standard error instead of individual standard errors of genes can be used, but is problematic when it is not reasonable to assume that all genes have the same variability (Cui and Churchill, 2003). Sometimes a modified version of the t-statistic such as SAM (Significance Analysis of Microarray), which uses a more stable estimate of standard error, is used (Tusher et al., 2001).

A nonparametric test such as the permutation test based on an N statistic is

sometimes more powerful than the ordinary t-test (Klebanov et al., 2004). Other nonparametric tests such as Wilcoxon's rank test, the Kolmogorov-Smirnov test, the Cramer-von Mises test and the Mann-Whitney test have also been used.

ANOVA modeling (Tusher et al., 2001) by means of bootstrapping (Kerr et al., 2002) has also been used to test hypotheses in microarray analyses.

Since microarray analyses involve testing many hypotheses simultaneously, methods of controlling type I error including FDR are employed (Benjamini and Hochberg, 1995; Dudoit et al., 2003).

1.4 The Outline of the Dissertation

This dissertation consists of 5 chapters. Chapter I provides an overview of the dissertation. Chapter II proposes methods for estimating the unknown density f. We consider two density estimation methods: a maximum likelihood unimodal density estimate and a kernel density estimate. Evidence is provided that the kernel density estimator is consistent. Methods of choosing an optimal number of clusters are also presented. In chapter III, we introduce location-scale invariant test statistics and a bootstrap algorithm to test hypotheses $H_{0i} : \beta_i = 0,\ i = 1, \ldots, G$. The t-statistic and an empirical likelihood ratio statistic are the tests employed. In chapter IV, we analyze a real data set and simulated data sets using methods introduced in previous chapters. The real data set comes from a comparative microarray experiment. We also simulate data sets from different mixture distributions to check the validity of the proposed methods. Finally we summarize conclusions and suggest future studies in chapter V.

CHAPTER II

DENSITY ESTIMATION BASED ON CLUSTERING

In the previous chapter, we described a statistical testing problem involving HDLSS data and explained that ordinary t-tests, sign tests and traditional bootstrapping may fail to be valid and powerful due to the small sample sizes for each subject. We introduce a mixture model in which data are pooled from all subjects to estimate a density f that determines the sampling distribution of each test statistic in the study.

In this chapter we describe methods for estimating f. These methods make use of clustering algorithms, a unimodal density estimator of Wegman (1969) and kernel density estimators.

2.1 Statistical Model

Our most general model is as follows. The jth measurement on the ith subject is L_{ij}, which satisfies

$$L_{ij} = \beta_i + \gamma_i \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G, \tag{2.1}$$

where all ϵ_{ij}'s are i.i.d. from an unknown density f. It is assumed that f has mean 0 and variance 1 and is unimodal.

The moment assumptions are not restrictive since β_i and γ_i are the mean and standard deviation, respectively, of L_{ij}. The unimodality assumption is to mitigate identifiability problems in a subsequent mixture model.

Our interest is in testing each of the null hypotheses $H_0 : \beta_i = 0$ for $i = 1, ..., G$, where G is the number of subjects in the data. We use test statistics with distributions

that are invariant to β_i and γ_i, and hence the distributions are completely determined by f. Given an estimate $\hat{f}$ of f, we may approximate the distribution of such test statistics by drawing bootstrap samples from $\hat{f}$.

Each subject belongs to one and only one of C clusters. All measurements from the same subject have to be from the same cluster. The cluster means are $\mu_1, \ldots, \mu_C$, and the standard deviations are $\sigma_1, \ldots, \sigma_C$. The model may then be written

$$L_{ij} = \mu_{k(i)} + \sigma_{k(i)}\epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G, \tag{2.2}$$

where $k(i)$ is the cluster to which subject i belongs. The model (2.2) is used only to get estimates of f. We do not impose equality of β_i's when we finally test each of $H_0 : \beta_i = 0$.

A measurement for a randomly selected subject from the model (2.2) has the mixture density

$$f_{\text{mixture}}(x) = \sum_{k=1}^{C} \frac{w_k}{\sigma_k} f\left(\frac{x - \mu_k}{\sigma_k}\right), \tag{2.3}$$

where w_k is the proportion of all subjects in the population that fall into cluster k, $k = 1, \ldots, C$, and the other parameters are the same as in (2.2). The conditional density of L_{ij} given that the ith subject belongs to group k is defined as

$$f(x|i \in \text{Cluster } k) = \frac{1}{\sigma_k} f\left(\frac{x - \mu_k}{\sigma_k}\right).$$

The following problems arise in estimating f.

(1) All parameters of mixture model must be estimated, since f cannot be estimated without knowledge of $w_1, \ldots, w_C$, $\mu_1, \ldots, \mu_C$ and $\sigma_1, \ldots, \sigma_C$.

(2) We do not know C, and so must treat it as unknown parameter.

Methods of estimating C will be discussed in section 2.3. In section 2.2, we just pretend that C is known, or set C to be $\tilde{C}$, and then f in (2.3) is estimated by using

either a maximum likelihood unimodal density estimator or a kernel density estimator for the given C.

The next section describes two density estimation methods.

2.2 Two Density Estimation Methods

When we estimate f by using either a maximum likelihood unimodal or kernel density estimator for a given $\tilde{C}$, f cannot be estimated without knowing $w_1, \ldots, w_{\tilde{C}}$, $\mu_1, \ldots, \mu_{\tilde{C}}$ and $\sigma_1, \ldots, \sigma_{\tilde{C}}$.

As we described before, we need to get estimates of parameters and specify which cluster each subject is from in the unimodal density estimation scheme. Clustering methods including agglomerative and K-means clustering may be applied to get $\hat{\mu}_1, \ldots, \hat{\mu}_{\tilde{C}}$ and $\hat{\sigma}_1, \ldots, \hat{\sigma}_{\tilde{C}}$ and to assign each subject to one and only one cluster.

A modified kernel density estimator of f for a given $\tilde{C}$ has the form

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i}^{G} \sum_{k=1}^{\tilde{C}} \hat{\alpha}_{ki} \sum_{j=1}^{n_i} K\left(\frac{x - e_{ij(k)}}{h}\right), \tag{2.4}$$

where all quantities are defined as in chapter I. In this form of the density estimator, we need to get $\hat{\alpha}_{ki}$'s, $\hat{\mu}_1, \ldots, \hat{\mu}_{\tilde{C}}$ and $\hat{\sigma}_1, \ldots, \hat{\sigma}_{\tilde{C}}$, and agglomerative or K-means clustering provides these estimates.

We will discuss clustering methods that lead to estimates of parameters in the next subsection.

2.2.1 Initial Estimates

The nonparametric density estimation methods proposed in this chapter require initial estimates of the components of the mixture model (2.3). Three types of initial estimates are considered: agglomerative clustering, K-means clustering, and Gaussian mixture model estimates (Fraley and Raftery, 2002).

2.2.1.1 Agglomerative Clustering

We consider two forms of agglomerative clustering (Fraley and Raftery, 2002): Gaussian-type and kernel estimation-type.

At each step of the former type of algorithm, it is assumed that observations within any given cluster are i.i.d. normal.

- At first, all subjects are considered as distinct clusters.
- Consider all $\binom{G}{2}$ ways of combining two subjects into one cluster, with the remaining $G-2$ subjects treated as distinct clusters, The mean and variance within each cluster are estimated by maximum likelihood. The two subjects that together maximize the likelihood are made into a single cluster.
- Combine into one cluster the two clusters that maximize the likelihood. After this step there are $G-1$ clusters.
- Repeat the previous 2 steps until all subjects are in the same cluster.

In another agglomerative clustering algorithm, the Gaussian model is replaced by a kernel density estimator. Before describing the algorithm, let A denote the set of indices for an arbitrary collection of subjects. Then

$$\hat{f}_A(x) = \frac{1}{N_A h_A} \sum_{i \in A} \sum_{j=1}^{n_i} K\left(\frac{x - L_{ij}}{h_A}\right),$$

where $N_A = \sum_{i \in A} n_i$, $h_A = \hat{\sigma}_A {N_A}^{-1/5}$ and $\hat{\sigma}_A$ is the sample standard deviation for the observations in A. The likelihood for this group of subjects is then

$$L_A = \prod_{i \in A} \prod_{j=1}^{n_i} \hat{f}_A(L_{ij}).$$

The overall likelihood for collections, or clusters, $A_1, \ldots, A_k$ $(k \leq G)$ is $\prod_{i=1}^{k} L_{A_i}$.

The steps of this algorithm are as follows:

- At first, all subjects are considered as distinct clusters.
- Combine into one cluster the two clusters that maximize the overall likelihood.
- Repeat the previous step until we have only one cluster.

2.2.1.2 K-means Clustering

A less computationally intensive method is K-means clustering (Hartigan, 1975). This algorithm seeks to minimize the sum of within cluster variances. An exhaustive search for the minimum is not feasible. The algorithm of Hartigan and Wong (1979) seeks local optima, i.e., solutions such that no movement of a data value from one cluster to another reduces the within cluster sum of squares.

After applying one of the clustering methods just described, parameter estimates for a given $\tilde{C}$ are obtained as follows.

$$\hat{\mu}_k = \sum_{i \in \text{Cluster } k} \sum_{j=1}^{n_i} \frac{L_{ij}}{N_k},$$

$$\hat{\sigma}_k = \frac{\left(\sum_{i \in \text{Cluster } k} \sum_{j=1}^{n_i} (L_{ij} - \hat{\mu}_k)^2 \right)^{\frac{1}{2}}}{\sqrt{N_k - 1}}$$

and

$$\hat{w}_k = \frac{N_k}{N},$$

where N is the total number or measurements, and N_k is the number of measurements within cluster k. Information about which subject is from which cluster is obtained as well. Then we are ready to apply density estimation methods.

2.2.1.3 Normal Mixture Models

Another method of getting initial estimates is fitting a Gaussian mixture model. The EM (Expectation Maximization) procedure of Fraley and Raftery (2002) for

fitting such a model is summarized in this subsection. Under the assumption that f is a standard normal density we apply the EM algorithm (Fraley and Raftery, 2002) to get initial estimates of proportions, means and variances of clusters for density estimation methods. When it is assumed that $\boldsymbol{L} = \{L_{ij} : j = 1, \ldots, n_i,\ i = 1, \ldots, G\}$ is a random sample from a mixture of normals, the likelihood function is

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w}|\boldsymbol{L}) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} \sum_{k=1}^{C} \frac{w_k}{\sigma_k} \phi \left(\frac{L_{ij} - \mu_k}{\sigma_k} \right),$$

where G is the number of subjects, n_i is the number of measurements from the ith subject, C is the number of components, and ϕ is the standard normal density.

The EM algorithm is a method of computing maximum likelihood estimators when the data consist of observed and unobserved parts. As initial values, one can use means and standard deviations from Gaussian agglomerative clustering.

Define $\boldsymbol{X}$ to be the complete data $(\boldsymbol{L}, \boldsymbol{\alpha})$ for subject i, where $\boldsymbol{L}_i = (L_{i1}, \ldots, L_{in_i})$, $i = 1, \ldots, G$, $\boldsymbol{\alpha}_i = (\alpha_{1i}, \ldots, \alpha_{Ci})$ and

$$\alpha_{ki} = \begin{cases} 1 & \text{if } \boldsymbol{L}_i \text{ belongs to Cluster } k \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_G$ are unobserved in our setting. Assume that C is known to be $\tilde{C}$. We will assume that the $\boldsymbol{\alpha}_i$'s are i.i.d. from a multinomial distribution of $\tilde{C}$ categories with probabilities $(w_1, \ldots, w_{\tilde{C}})$, in which case the probability function of $\boldsymbol{\alpha}_i$ is

$$f(\boldsymbol{\alpha}_i) = \prod_{k=1}^{\tilde{C}} {w_k}^{\alpha_{ki}}.$$

The density of the observed $\boldsymbol{L}_i$ given $\boldsymbol{\alpha}_i$ is

$$\prod_{k=1}^{\tilde{C}} \left[\prod_{j=1}^{n_i} \frac{1}{\sigma_k} \phi \left(\frac{L_{ij} - \mu_k}{\sigma_k} \right) \right]^{\alpha_{ki}}.$$

Therefore, the complete-data log-likelihood is

$$l(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{G} \sum_{k=1}^{\tilde{C}} \alpha_{ki} \left[\log(w_k) + \sum_{j=1}^{n_i} \log \left(\frac{1}{\sigma_k} \phi \left(\frac{L_{ij} - \mu_k}{\sigma_k} \right) \right) \right]$$

where $\boldsymbol{x}$ is a matrix of $\{\boldsymbol{L}_i,\ \boldsymbol{\alpha}_i,\ i = 1, \ldots, G\}$. A description of the EM algorithm for the normal mixture model is now given. The algorithm is iterative, and the quantity r in our description denotes iteration number. Agglomerative clustering provides parameter estimates for $r = 0$.

- The E Step of the EM algorithm provides an estimate of $P(i \in \text{Cluster } k|\boldsymbol{L}_i)$:

$$\hat{\alpha}_{ki}^r = \frac{\hat{w_k}^r \prod_{j=1}^{n_i} \frac{1}{\hat{\sigma_k}^r} \phi \left(\frac{L_{ij} - \hat{\mu_k}^r}{\hat{\sigma_k}^r} \right)}{\sum_{l=1}^{\tilde{C}} \hat{w_l}^r \prod_{j=1}^{n_i} \frac{1}{\hat{\sigma_l}^r} \phi \left(\frac{L_{ij} - \hat{\mu_l}^r}{\hat{\sigma_l}^r} \right)}.$$

- The M Step of the EM algorithm finds estimates of $\boldsymbol{\mu}, \boldsymbol{\sigma}$ and $\boldsymbol{w}$ by maximizing the complete-data log-likelihood. The forms of the MLEs (Fraley and Raftery, 2002) are

$$\hat{w}_k^{r+1} = \frac{\sum_{i=1}^{G} \hat{\alpha}_{ki}^r}{G},$$

$$\hat{\mu}_k^{r+1} = \frac{\sum_{i=1}^{G} \sum_{j=1}^{n_i} \hat{\alpha}_{ki}^r L_{ij}}{\sum_{i=1}^{G} n_i \hat{\alpha}_{ki}^r},$$

and

$$\hat{\sigma}_k^{r+1} = \sqrt{\frac{\sum_{i=1}^{G} \sum_{j=1}^{n_i} \hat{\alpha}_{ki}^r (L_{ij} - \hat{\mu}_k^{r+1})^2}{\sum_{i=1}^{G} n_i \hat{\alpha}_{ki}^r}}.$$

- Repeat the E and M steps until the MLEs of the proportions, the means and the standard deviations converge.

2.2.2 Nonparametric Density Estimation Methods

Once parameter estimates for a given $\tilde{C}$ and clustering information are obtained, density estimation methods are applied to get an estimate of f. In this section we describe two types of methodology for estimating f. Each of these can take on various forms, but we only describe the fundamental difference between the two.

The errors, $\epsilon_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, G$, are i.i.d. as f, and so if they could be observed, they could be used in any of several standard density estimation methods to obtain an estimate of f.

2.2.2.1 Cluster-Based Estimation

The cluster-based estimation method operates by approximating the errors. For a given number $\tilde{C}$ of clusters, divide the G subjects into $\tilde{C}$ clusters using some clustering algorithm. In principle, the algorithm could be arbitrary. It might be, for example, agglomerative or K-means.

The clustering algorithm provides estimates $\hat{\mu}_k$ and $\hat{\sigma}_k$ of μ_k and σ_k for cluster k, $k = 1, \ldots, \tilde{C}$. Now residuals are computed as follows:

$$e_{ij} = \frac{L_{ij} - \hat{\mu}_{k(i)}}{\hat{\sigma}_{k(i)}}, \quad i = 1, \ldots, n_j, \quad i = 1, \ldots, G, \tag{2.5}$$

where $k(i)$ denotes the cluster into which subject i has been assigned. If the assignment of subjects to clusters approximates well true cluster membership, then, as a whole, these residuals will be reasonable approximations to the error terms $\epsilon_{ij},\ j = 1, \ldots, n_j,\ i = 1, \ldots, G$.

One method of estimating f from the residuals (2.5) is to use an ordinary kernel estimator, which would take the form

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{G} \sum_{j=1}^{n_j} K\left(\frac{x - e_{ij}}{h}\right). \tag{2.6}$$

Another means of estimating f is to use the maximum likelihood unimodal density estimator investigated by Wegman (1969 ; 1970) and Meyer (2001) . In principle, it is possible to obtain a near maximum likelihood estimate of the mixture model, f_{mixture}, assuming a given value $\tilde{C}$ for C and that f is unimodal. We now describe an algorithm for doing so.

Let $\boldsymbol{A}$ denote a given way of grouping the G subjects into $\tilde{C}$ clusters. For the grouping $\boldsymbol{A}$, let $\hat{\mu}_k(\boldsymbol{A})$, $\hat{\sigma}_k(\boldsymbol{A})$, and $N_k(\boldsymbol{A})$ denote the sample mean, sample standard deviation and number of observations in cluster k, $k = 1, \ldots, \tilde{C}$. Standardized data may now be computed exactly as in (2.5). Finally, the maximum likelihood unimodal estimator (Wegman, 1969), call it $\hat{f}(\cdot|\boldsymbol{A})$, may be computed from the standardized data.

Defining

$$f_{\text{mixture}}(x|\boldsymbol{A}) = \sum_{k=1}^{\tilde{C}} \frac{N_k(\mathbf{A})}{N} \cdot \frac{1}{\hat{\sigma}_k(\boldsymbol{A})} \hat{f}\left(\frac{x - \hat{\mu}_k(\mathbf{A})}{\hat{\sigma}_k(\boldsymbol{A})}|\boldsymbol{A}\right),$$

we may compute the likelihood

$$L(\mathbf{A}) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} f_{\text{mixture}}(L_{ij}|\boldsymbol{A}).$$

At this point, one may search over all groupings $\boldsymbol{A}$ to find one that maximizes $L(\mathbf{A})$. The problem of searching for an optimal grouping is not dealt with in this dissertation.

We now describe how the unimodal density estimator is computed. We assume the mode is known to be 0. (Meyer (2001) describes how to deal with the case of an unknown mode.)

By using the result of one of the clustering methods, we standardize the data as in (2.5). We pretend the e_{ij}'s are a random sample from f, and denote the sorted e_{ij}'s by $e_{(1)} < e_{(2)} \ldots < e_{(N)}$. Following Meyer (2001), when 0 is in the interval $(e_{(m-1)}, e_{(m)}]$,

upper sets $U = \{u_1, \ldots, m, \ldots, u_2\}$ and lower sets $L = \{2, \ldots, l_1\} \cup \{l_2, \ldots, N\}$ are defined, where $l_1 < m$ and $l_2 > m$ or $l_1 = l_2 = m$.

The unimodal density estimator is defined by $\tilde{f}(x) = \tilde{\theta}_j$ for $x \in (e_{(j-1)}, e_{(j)}]$, $j = 2, \ldots, N$, where

$$\tilde{\theta_j} = \max_{U: j \in U} \min_{L: j \in L} \frac{(l_1 - u_1 + 1)_+/N + (u_2 - l_2 + 1)_+/N}{(e_{(l_1)} - e_{(u_1 - 1)})_+ + (e_{(u_2)} - e_{(l_2 - 1)})_+}$$

and $(x)_+ = \max(0, x)$. The unimodal density estimate in the interval containing the mode is

$$\tilde{\theta_m} = \max_{u_1 \le m \le u_2} \frac{(u_2 - u_1 + 1)/N}{e_{(u_2)} - e_{(u_1 - 1)}}.$$

The main drawback of the cluster-based method is that it relies on the initial clustering scheme. One way of addressing this problem is to iterate the method. The initial estimate $\hat{f}$ of f could be used to define a new clustering algorithm. After the data have been grouped into $\tilde{C}$ clusters, the scheme described above could be carried out again with the new clustering. One possibility for the second-stage clustering would be agglomerative with the standard normal density replaced by $\hat{f}$.

2.2.2.2 A Modified Kernel Density Estimator

Our second main proposal for estimating f is to use a modified sort of kernel density estimator. Proportions, means and variances from a clustering algorithm are used as initial values, and then kernel density estimation is used to estimate f. We will discuss estimating the optimal number of clusters in the next section, but here C is assumed to be given.

Suppose that all the parameters of the mixture model, f_{mixture}, were known, including C and f. Let $\boldsymbol{L}_i = (L_{i1}, \ldots, L_{in_i})$ denote observations from a subject i

whose cluster membership is unknown, and define

$$g_k(\boldsymbol{L}_i) = w_k \prod_{k=1}^{n_i} \frac{1}{\sigma_k} f\left(\frac{L_{ij} - \mu_k}{\sigma_k}\right), \quad k = 1, \ldots, C.$$

The conditional probability that subject i comes from cluster k is defined as

$$p(k|\boldsymbol{L}_i) = \alpha_{ki} = \frac{g_k(\boldsymbol{L}_i)}{\sum_{l=1}^{C} g_l(\boldsymbol{L}_i)}.$$

Now define standardized observations

$$e_{ij(k)} = \frac{L_{ij} - \hat{\mu}_k}{\hat{\sigma}_k}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G, \quad k = 1, \ldots, C. \tag{2.7}$$

Note that $e_{ij(k)}$ is the correct standardization of observation L_{ij} if i comes from cluster k.

Now, a kernel estimate of f may be defined as follows:

$$\begin{aligned} \tilde{f}_h(x) &= \frac{1}{Nh} \sum_{i=1}^{G} \sum_{k=1}^{C} \alpha_{ki} \sum_{j=1}^{n_i} K\left(\frac{x - e_{ij(k)}}{h}\right) \\ &= \sum_{i=1}^{G} \frac{n_i}{N} \sum_{k=1}^{C} \alpha_{ki} \hat{f}_h(x|i,k), \end{aligned}$$

where

$$\hat{f}_h(x|i,k) = \frac{1}{n_i h} \sum_{j=1}^{n_i} K\left(\frac{x - e_{ij(k)}}{h}\right).$$

Note that $\hat{f}_h(\cdot|i,k)$ is an ordinary kernel estimate using standardized data from subject i and the assumption that i is from cluster k. The contribution of each subject to $\tilde{f}$ is a weighted average of $\hat{f}(\cdot|i,k)$, $k = 1, \ldots, C$, with each weight equaling the probability that the subject belongs to a given cluster.

Of course, in practice $\tilde{f}_h$ is not even an estimator since it depends upon unknown parameters and the very function f, that is to be estimated. To circumvent this problem, we propose that a Gaussian mixture model be fitted to the data by using

some candidate $\tilde{C}$ for C. Likewise, estimates of μ_k and σ_k are used in defining standardized observations as in (2.7). Now, an estimate of the form $\tilde{f}_h$ may be computed in an obvious way using quantities so-defined.

Define the density estimate of f at the first iteration to be

$$\hat{f}_{h,1}(x) = \frac{1}{Nh} \sum_{m=1}^{G} \sum_{n=1}^{n_m} \sum_{l=1}^{\tilde{C}} \hat{\alpha}_{lm}^0 K\left(\frac{x - e_{mn(l)}^0}{h}\right).$$

The quantities $\hat{\alpha}_{lm}^0$ and $e_{mn(l)}^0$ depend upon $\hat{\boldsymbol{w}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\sigma}}$ and $\hat{f}_{h,0}$, and $\hat{f}_{h,0}$ is an initial estimate of f. If we use a Gaussian mixture model to obtain initial estimates, then $\hat{f}_{h,0}$ is equivalent to ϕ. Otherwise

$$\hat{f}_{h,0}(x) = \sum_{l=1}^{\tilde{C}} \hat{w}_l \frac{1}{N_l h_l} \sum_{m \in \text{Cluster } l} \sum_{n=1}^{n_m} K\left(\frac{x - e_{mn}^0}{h_l}\right),$$

where K is the Gaussian kernel, i.e.,

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

N_l is the total number of measurements within cluster l,

$$e_{mn}^0 = \frac{L_{mn} - \hat{\mu}_{l(m)}^0}{\hat{\sigma}_{l(m)}^0},$$

and h_l is the bandwidth, which we take to be $h_l = {N_l}^{-1/5}$. New parameter estimates are obtained by maximizing

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w}) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} \sum_{k=1}^{\tilde{C}} \frac{w_k}{\sigma_k} \hat{f}_{h,1}\left(\frac{L_{ij} - \mu_k}{\sigma_k}\right). \tag{2.8}$$

These new estimates lead to $\hat{f}_{h,2}(x)$ by replacing (in $\hat{f}_{h,1}$) $\hat{f}_{h,0}$ by $\hat{f}_{h,1}$ and the previous estimates of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{w}$ by new ones. This process may then be iterated. In general, a kernel density estimate of f at the r-th iteration ($r = 1, 2, \ldots$) has the form

$$\hat{f}_{h,r}(x) = \frac{1}{Nh} \sum_{m=1}^{G} \sum_{n=1}^{n_m} \sum_{l=1}^{\tilde{C}} \hat{\alpha}_{lm}^{r-1} K\left(\frac{x - {e_{mn(l)}}^{r-1}}{h}\right). \tag{2.9}$$

It is not difficult to show that (2.9) is a density since it always takes positive values and integrates to 1, as shown below:

$$\begin{aligned}
\int \hat{f}_{h,r}(x)dx &= \int \frac{1}{Nh}\sum_{m=1}^{G}\sum_{n=1}^{n_m}\sum_{l=1}^{\tilde{C}}\hat{\alpha}_{lm}^{r-1}K\left(\frac{x-e_{mn(l)}^{r-1}}{h}\right)dx \\
&= \frac{1}{Nh}\sum_{m=1}^{G}\sum_{n=1}^{n_m}\sum_{l=1}^{\tilde{C}}\hat{\alpha}_{lm}^{r-1}\int K\left(\frac{x-e_{mn(l)}^{r-1}}{h}\right)dx \\
&= \frac{h}{Nh}\sum_{m=1}^{G}\sum_{n=1}^{n_m}\sum_{l=1}^{\tilde{C}}\hat{\alpha}_{lm}^{r-1}\int K(u)du \\
&= \frac{1}{N}\sum_{m=1}^{G}\sum_{n=1}^{n_m}\sum_{l=1}^{\tilde{C}}\hat{\alpha}_{lm}^{r-1} \\
&= \sum_{m=1}^{G}\sum_{l=1}^{\tilde{C}}\hat{\alpha}_{lm}^{r-1}\frac{n_m}{N} \\
&= \frac{1}{N}\sum_{m=1}^{G}n_m,
\end{aligned}$$

since the $\hat{\alpha}_{lm}^{r-1}$ sum to 1 for each m. The last quantity is 1 since $N = \sum_{m=1}^{G} n_m$.

Two important issues in kernel density estimation are kernel selection and bandwidth selection. We will use a Gaussian kernel since it is well known that kernel density estimators are usually not sensitive to choice of K, at least within a reasonable class of kernels that are densities. The selection of h is more crucial. We use a bandwidth that would be asymptotically optimal mean integrated squared error were the data a random sample from a normal density. This bandwidth has the form

$$h_n = 1.06\sigma n^{-1/5},$$

for a sample size of n when the population standard deviation is σ and a Gaussian kernel is used. In our setting where f has $\sigma = 1$, the bandwidth takes the even simpler form

$$h_N = 1.06N^{-1/5}.$$

This bandwidth is clearly not optimal in our setting since the kernel estimator is not of standard form and f is not necessarily normal. However, h_N has the virtue of simplicity and stability in comparison to a procedure that attempts to estimate the optimal bandwidth.

There are no closed form MLEs for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ in (2.8), so we consider three different methods for estimating $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.

- For iteration r, one can use Newton-Raphson with initial estimates of $\boldsymbol{w}, \boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ equal to the estimates from iteration $r-1$.

- R-step MLE : Often the computing time needed for convergence of Newton-Raphson is prohibitive, and hence we could stop after some pre-specified number, R, of iterations.

- Pseudo MLE : We can also use estimates having the same form as those in a Gaussian mixture model, i.e.,

$$\hat{\mu}_k^r = \frac{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\hat{\alpha}_{ki}^{r-1}L_{ij}}{\sum_{i=1}^{G} n_i\hat{\alpha}_{ki}^{r-1}},$$

$$\hat{\sigma}_k^r = \sqrt{\frac{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\hat{\alpha}_{ki}^{r-1}(L_{ij}-\hat{\mu}_k^r)^2}{\sum_{i=1}^{G} n_i\hat{\alpha}_{ki}^{r-1}}}.$$

- For a given $\tilde{C}$ we iterate the E and M steps by updating the empirical density, which is $\hat{f}_{h,r}$, until $\hat{w}_k,\ \hat{\mu}_k$ and $\hat{\sigma}_k$ converge.

An estimate of f corresponding to a given number of clusters, $\tilde{C}$, is obtained by using one of the two main density estimation methods discussed in this section, and in the next section we will discuss how to estimate an optimal number of clusters.

2.3 Estimating the Number of Clusters, C

Now we discuss how to estimate C, the number of components in the mixture model. There are several ways to choose an optimal value of C.

Using the maximum likelihood principle, that chooses C to maximize likelihood, leads to choosing the largest C considered, so we need a modified criterion.

One possible criterion is AIC (Akaike Information Criterion), which in our setting is

$$AIC(C) = 2\log \hat{L}_C - 2(3C-1),$$

where $\hat{L}_C$ is the maximized likelihood for a given C. We choose not to use AIC since it tends to overestimate the true dimension of the model.

Another criterion is BIC (Bayes Information Criterion), which also subtracts a penalty for model dimension from the log likelihood. In our setting, BIC takes the form

$$BIC(C) = 2\log \hat{L}_C - (3C-1)\log N, \quad C = 1, 2, \ldots,$$

where $\hat{L}_C = \prod_{i=1}^{G}\prod_{j=1}^{n_i} \hat{f}_h(L_{ij})$ and $\hat{f}_h$ is as in (2.6). We choose C to maximize $BIC(C)$, and use the estimate of f corresponding to the maximizer. Because of its larger penalty term, BIC usually chooses a smaller C than does AIC.

The third criterion is based on the *weighted mean of within cluster variance estimates (WMCV)*, which has the form

$$WMCV(C) = \sum_{k=1}^{C} \frac{\hat{N}_k}{G} \hat{\sigma_k}^2,$$

where C is the number of clusters, $\hat{N}_k$ is the number of subjects within cluster k, G is the total number of subjects and $\hat{\sigma_k}^2$ is the kth cluster variance estimate, i.e.,

$$\hat{\sigma}_k^2 = \frac{1}{\hat{N}_k - 1} \sum_{i \in \text{Cluster } k} \sum_{j=1}^{n_i} (L_{ij} - \bar{L_{i.}})^2.$$

Clusters are obtained by applying a clustering method such as agglomerative or K-means clustering.

The mean of all G within subject variances is

$$\frac{1}{G}\sum_{i=1}^{G}{\gamma_i}^2. \tag{2.10}$$

Under our mixture model we have

$$\frac{1}{G}\sum_{i=1}^{G}{\gamma_i}^2 = \frac{1}{G}\sum_{k=1}^{C}N_k{\sigma_k}^2, \tag{2.11}$$

where N_k is the number of subjects within the cluster k. Formula (2.11) shows that the mean of the G within subject variances should be close to $WMCV$ when the latter is computed with the true value of C and cluster assignments that well-approximate the truth.

Expression (2.10) may be estimated by

$$MSV = \frac{1}{G}\sum_{i=1}^{G}\hat{\gamma}_i^2 = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{n_i-1}\sum_{j=1}^{n_i}(L_{ij}-\bar{L}_{i\cdot})^2.$$

By a theorem of Kolmogorov (Serfling, 1980) this estimator is strongly consistent for $\sum_{k=1}^{C}w_k\sigma_k^2$ as $G\to\infty$ whenever f has finite fourth moment.

Let C_{max} be an upper bound on the number of clusters considered. Then we may estimate the true number of clusters by the value of C that minimizes $|WMCV(C)-MSV|$, $C=1,\ldots,C_{max}$.

We will use either WMCV or BIC to estimate the true number of clusters. Once the optimal number of clusters has been decided, we apply density estimation methods as already discussed in the previous section to get an estimate of f.

2.4 Investigating Consistency of the Kernel Density Estimate

We continue to assume that the true number of clusters is known to be C. To simplify notation, we assume throughout this section that there is only one observa-

tion per subject. Now define

$$p_k(x|d,(\boldsymbol{m},\boldsymbol{s},\boldsymbol{a})) = \frac{a_k \frac{1}{s_k} d\left(\frac{x-m_k}{s_k}\right)}{\sum_{l=1}^{C} a_l \frac{1}{s_l} d\left(\frac{x-m_l}{s_l}\right)},$$

where d is an arbitrary density with mean 0 and variance 1, and $(\boldsymbol{m},\boldsymbol{s},\boldsymbol{a})$ are arbitrary choices for the mixture model parameters.

Under general conditions, the parameter estimators of $(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w})$ obtained by fitting a Gaussian mixture model will be consistent for certain quantities, call them $(\boldsymbol{\mu}_0,\boldsymbol{\sigma}_0,\boldsymbol{w}_0)$, that are not necessarily equal to $(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w})$. These are the parameters that produce a best normal mixture approximation of f_{mixture} in the sense of Kullback-Leibler information divergence. Assuming that this consistency holds, the first stage kernel estimator will be consistent for

$$\lim_{h\to 0} E\left[\frac{1}{Gh}\sum_{i=1}^{G}\sum_{k=1}^{C} p_k(L_i|\phi,(\boldsymbol{\mu}_0,\boldsymbol{\sigma}_0,\boldsymbol{w}_0)) K\left(\frac{\sigma_{0k}x - L_i + \mu_{0k}}{h\sigma_{0k}}\right)\right]$$

as $Gh \to \infty$ and $h \to 0$. The expectation in the last expression is

$$\sum_{k=1}^{C} \frac{1}{h}\int_{-\infty}^{\infty} p_k(y|\phi,(\boldsymbol{\mu}_0,\boldsymbol{\sigma}_0,\boldsymbol{w}_0)) K\left(\frac{\sigma_{0k}x - y + \mu_{0k}}{h\sigma_{0k}}\right) f_{\text{mixture}}(y)\, dy.$$

Make the change of variable $z = (\sigma_{0k}x - y + \mu_{0k})/(\sigma_{0k}h)$ in each integral, $k = 1,\ldots,C$, and the last expression becomes

$$\sum_{k=1}^{C} \sigma_{0k}\int_{-\infty}^{\infty} p_k(\sigma_{0k}x + \mu_{0k} - \sigma_{0k}hz|\phi,(\boldsymbol{\mu}_0,\boldsymbol{\sigma}_0,\boldsymbol{w}_0)) K(z) \\ f_{\text{mixture}}(\sigma_{0k}x + \mu_{0k} - \sigma_{0k}hz)\, dz.$$

Assuming that f_{mixture} is continuous, the last expression tends to the density $f_0(x)$ $\equiv f(x|\phi,(\boldsymbol{\mu}_0,\boldsymbol{\sigma}_0,\boldsymbol{w}_0))$ as $h \to 0$, where, for an arbitrary density d and parameter vector $(\boldsymbol{m},\boldsymbol{s},\boldsymbol{a})$,

$$f(x|d,(\boldsymbol{m},\boldsymbol{s},\boldsymbol{a})) = \sum_{k=1}^{C} s_k p_k(s_k x + m_k|d,(\boldsymbol{m},\boldsymbol{s},\boldsymbol{a})) f_{\text{mixture}}(s_k x + m_k). \qquad (2.12)$$

It is easy to verify that $f(x|f,(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w}))\equiv f$. This seemingly trivial observation is nonetheless crucial since if it were not true, iterating the estimation scheme would have little (if any) chance of success. The second-stage estimators of $(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w})$ (as defined in (2.14)) will, generally speaking, be consistent for parameters $(\boldsymbol{\mu}_1,\boldsymbol{\sigma}_1,\boldsymbol{w}_1)$ that produce the best Kullback-Leibler approximation of f_{mixture} among all approximations of the form

$$\sum_{k=1}^{C} a_k \frac{1}{s_k} f_0\left(\frac{x-m_k}{s_k}\right).$$

The second-stage kernel estimator is consistent for $f_1(x)\equiv f(x|f_0,(\boldsymbol{\mu}_1,\boldsymbol{\sigma}_1,\boldsymbol{w}_1))$. This iteration scheme can be continued indefinitely, and the key question is whether or not $f_\infty(x)\equiv f(x)$. The answer to this question under very general conditions is beyond the scope of this dissertation. However, we will investigate the question numerically for a situation in which the parameters of the mixture model are assumed to be estimated consistently.

Define $f_0(x)\equiv f(x|\phi,(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w}))$ and

$$f_r(x)\equiv f(x|f_{r-1},(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w})),\quad r=1,2,\ldots. \tag{2.13}$$

Does f_r converge to f as $r\to\infty$? Before showing numerical results that address this question, we make some simple observations about f_r. Consider an "easy" situation in which the μ_j's are well-separated relative to the σ_k's and no w_k is exceptionally small. Then for $|x|<3$, we have $p_k(\sigma_k x+\mu_k|f_{r-1},(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w}))\approx 1$ for each k,

$$f_{\text{mixture}}(\sigma_k x+\mu_k)\approx w_k\frac{1}{\sigma_k}f(x)\quad\text{for each } k$$

and hence

$$f_r(x)\approx\sum_{k=1}^{C} w_k f(x)=f(x)\sum_{k=1}^{C} w_k=f(x).$$

This suggests that in the case of well-separated μ_k's, the kernel estimation scheme will produce a good estimate of f, at least if the mixture parameters are well-estimated.

Define $g \equiv f_{\text{mixture}}$, let $\hat{f}$ be any approximation for $f(x)$ that has mean 0 and variance 1, and define

$$\hat{g}(x) = \sum_{k=1}^{C} w_k \frac{1}{\sigma_k} \hat{f}\left(\frac{x-\mu_k}{\sigma_k}\right).$$

An updated approximation of f, as suggested by (2.12), is $f(x|\hat{f}, (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w}))$. We then have

$$\begin{aligned} f(x) \;-\; f(x|\hat{f}, (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w})) &= f(x) - \sum_{k=1}^{C} w_k \frac{\hat{f}(x)}{\hat{g}(\sigma_k x + \mu_k)} g(\sigma_k x + \mu_k) \\ &= f(x) - \hat{f}(x) \sum_{k=1}^{C} w_j \left[\frac{g(\sigma_k x + \mu_k) - \hat{g}(\sigma_k x + \mu_k) + \hat{g}(\sigma_k x + \mu_k)}{\hat{g}(\sigma_k x + \mu_k)}\right] \\ &= f(x) - \hat{f}(x) + \hat{f}(x) \sum_{k=1}^{C} w_k \left[\frac{\hat{g}(\sigma_k x + \mu_k) - g(\sigma_k x + \mu_k)}{\hat{g}(\sigma_k x + \mu_k)}\right] \\ &= f(x) - \hat{f}(x) + \hat{f}(x)\delta(x). \end{aligned} \tag{2.14}$$

So, the error in the updated approximation is equal to the old error, $f(x) - \hat{f}(x)$, plus the term $\hat{f}(x)\delta(x)$. In order for the new error to be smaller in magnitude than the old, it is necessary and sufficient that

(i) $\delta(x)$ be opposite in sign from $f(x) - \hat{f}(x)$, and

(ii) $\hat{f}(x)|\delta(x)| < 2|f(x) - \hat{f}(x)|$.

Expression (2.14) offers some hope that (i) may generally be true, since if $f(x) > \hat{f}(x)$ in, say, a neighborhood of 0, then the tendency will be for $\hat{g}(\sigma_k x + \mu_k) < g(\sigma_k x + \mu_k)$ for x in the same neighborhood. A sufficient condition for (ii) is

$$\sum_{k=1}^{C} w_k \frac{|\hat{g}(\sigma_k x + \mu_k) - g(\sigma_k x + \mu_k)|}{\hat{g}(\sigma_k x + \mu_k)} < 2\frac{|f(x) - \hat{f}(x)|}{\hat{f}(x)}. \tag{2.15}$$

If (i) is true, then (2.15) seems plausible in that it only requires the average relative error in the $\hat{g}(\sigma_k x + \mu_k)$'s to be less than twice the relative error of $\hat{f}$.

We now give numerical results that provide convincing evidence that f_r often

converges to f as $r \to \infty$. We investigate how close $f_r(x)$ is to $f(x)$ by means of example. We consider 3 different distributions for f, each one having mean 0 and variance 1. The densities are a Laplace

$$g(x) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|x|),$$

a gamma density,

$$g(x) = 2(x + \sqrt{2}) \exp(-\sqrt{2}(x + \sqrt{2}))I(x > -\sqrt{2}),$$

and a rescaled t-distribution with 3 degrees of freedom,

$$g(x) = \frac{2}{\pi} \frac{1}{(1 + x^2)^2}.$$

We take C, the number of clusters, to be 2, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$ and examine f_r for $w_1 = 0.5,\ 0.7,\ 0.9$ and $\mu_2 = 0.5,\ 1.6,\ 2.7,\ 3.8,\ 4.9,\ 6.0$. The plots of each mixture and the corresponding f_r's are given. In each case, the initial estimate for f is ϕ, and subsequent iterations are f_r, as defined by (2.13). After just a few iterations, in most cases f_r is very close to f. Plots in Figure 1 are the Laplace density with $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 0.5$, $\sigma_1 = 1$ and $\sigma_2 = 1$, and we see that f_r is almost indistinguishable from f after 5 iterations. The rest of the plots are in Appendices B through D.

The discussion to this point has focused on the effectiveness of the iterative kernel method in the absence of error in estimating $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w})$. In practice, of course, these parameters are unknown and must be estimated from the data. When the estimation process is iterated, we suggested in section 2.2.2.2 that the likelihood estimates of $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w})$ be computed on the assumption that the newest kernel estimate is the true f.

It is anticipated that normal mixture model MLEs of $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w})$ are consistent for some $(\boldsymbol{\mu}^0, \boldsymbol{\sigma}^0, \boldsymbol{w}^0)$, which is not necessarily equal to the true parameter vector unless

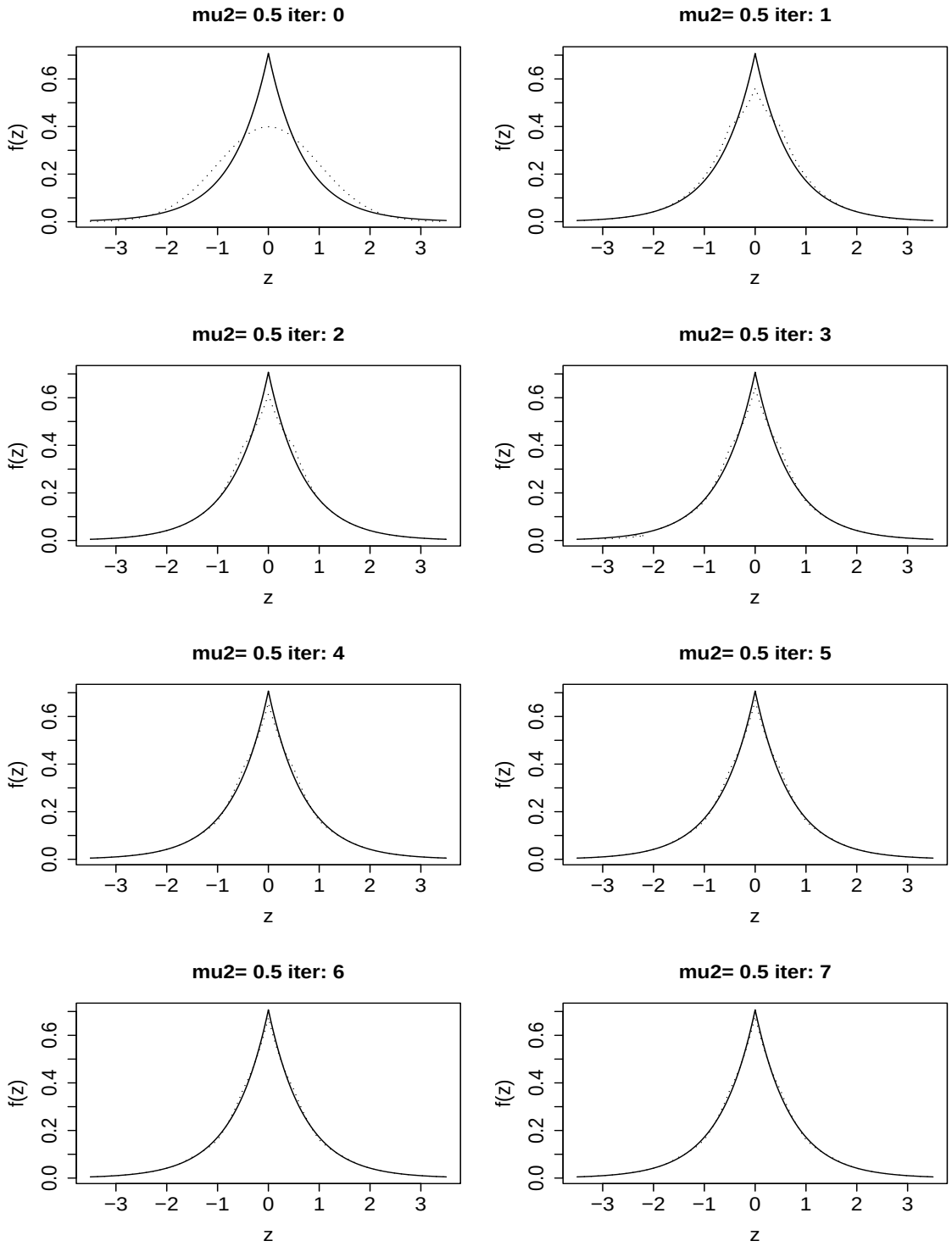

Figure 1: f_r for Laplace : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$

$f \equiv \phi$. The first iteration kernel estimate will then be consistent for

$$f_0(x) = \sum_{k=1}^{C} \sigma_k^0 \alpha_k^0 (\sigma_k^0 x + \mu_k^0) g(\sigma_k^0 x + \mu_k^0),$$

where

$$\alpha_k^0(y) = \frac{w_k^0 \frac{1}{\sigma_k^0} \phi(\frac{y-\mu_k^0}{\sigma_j^0})}{\sum_{l=1}^{C} w_l^0 \frac{1}{\sigma_l^0} \phi(\frac{y-\mu_l^0}{\sigma_l^0})}.$$

Maximum likelihood estimates of $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w})$ for a mixture model with $f \equiv f_0$ are anticipated to be consistent for some parameter vector $(\boldsymbol{\mu}^1, \boldsymbol{\sigma}^1, \boldsymbol{w}^1)$. The second stage kernel estimate is then consistent for

$$f_1(x) = \sum_{k=1}^{C} \sigma_k^1 \alpha_k^1 (\sigma_k^1 x + \mu_k^1) g(\sigma_k^1 x + \mu_k^1),$$

where α_k^1 is defined as was α_k^0 with $\boldsymbol{\mu}^0, \boldsymbol{\sigma}^0, \boldsymbol{w}^0$ replaced by $\boldsymbol{\mu}^1, \boldsymbol{\sigma}^1, \boldsymbol{w}^1$ and ϕ by f_0. This process may continue to be iterated, and the key question becomes under what conditions do $\boldsymbol{\mu}^r, \boldsymbol{\sigma}^r, \boldsymbol{w}^r$ and f_r converge to $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w})$ and f, respectively.

This is a very difficult mathematical question that is beyond the scope of this dissertation. However, our empirical results indicate that iterating past the normal mixture model usually produces better estimates than that model when the components are not Gaussian. Results to this effect will be given in chapter IV.

CHAPTER III

TEST STATISTICS AND BOOTSTRAP

Recall that our model is

$$L_{ij} = \beta_i + \gamma_i \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G,$$

where the ϵ_{ij}'s are i.i.d. from density f. We want to test the hypotheses $H_{0i} : \beta_i = 0,\ i = 1, \ldots, G$, where β_i is the population mean for the ith subject. In this chapter we discuss possible test statistics and describe our bootstrap procedure for approximating the distribution of a test statistic.

3.1 Commonly Used Tests

Commonly used tests of the hypothesis $H_0 : \beta_i = 0$ include the ordinary t-test, the sign test, the signed-rank test and bootstrap tests. For subject i the t-statistic has the form

$$T = \frac{\bar{L}_{i\cdot}}{s_i/\sqrt{n_i}},$$

where n_i is the number of measurements for the ith subject, and $\bar{L}_{i\cdot}$ and s_i are the sample mean and standard deviation, respectively, for subject i.

When $L_{i1}, \ldots, L_{in_i}$ are a random sample from $N(0, \gamma_i^2)$, the t-statistic follows the t distribution with $n_i - 1$ degrees of freedom. Even if each L_{ij} is not normally distributed, the central limit theorem guarantees that the t-statistic has an approximate standard normal distribution when $\beta_i = 0$ and n_i is large enough. However, we are interested in settings (as in microarray analyses) where n_i is quite small for every i. In such cases, the central limit theorem cannot be relied upon to protect against

nonnormality of the data.

Sign tests can be considered as an alternative to the t-test for small samples. The sign test only requires the assumption that $L_{i1}, \ldots, L_{in_i}$ are i.i.d. It is used to test the hypothesis $H_0 : \theta = 0$, where θ is the population median. The test statistic for a sign test of this hypothesis is

$$S = \sum_{j=1}^{n_i} I(L_{ij} > 0)$$

where $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. When $\theta = 0$, S follows the binomial distribution with number of trials n_i and success probability 0.5. The sign test deals effectively with nonnormality, but has poor power for very small sample sizes.

The signed-rank test is another alternative to the t-test. It tends to be more powerful than the sign test, but also requires symmetry of f for validity.

Bootstrapping is another way of dealing with nonnormality. It determines the distribution of the test statistic when sampling repeatedly from the empirical distribution. However, if n_i is only 3 or 4, the empirical distribution will usually be a poor estimate of the population distribution, and hence the bootstrap sampling distribution may be a poor estimate of the true sampling distribution.

3.2 Location and Scale Invariant Tests

Our primary interest is in test statistics whose distributions are invariant to changes in scale under H_0. If this is true, then the distribution of the test statistic is completely determined by f.

The t-statistic is obviously scale invariant under H_0, since in that case

$$\begin{aligned} T &= \frac{\bar{L}_{i\cdot}}{s_i/\sqrt{n_i}} \\ &= \frac{\gamma_i \bar{\epsilon}_{i\cdot}}{\gamma_i s_\epsilon/\sqrt{n_i}} \\ &= \frac{\bar{\epsilon_{i\cdot}}}{s_\epsilon/\sqrt{n_i}}, \end{aligned}$$

where $s_\epsilon^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2$. Obviously, T is free of β_i and γ_i under the null hypothesis, and the distribution of T is completely determined by f (and n_i). The bootstrapping algorithm to be explained in section 3.3 will be applied to approximate the null distribution of a t-statistic.

If f were known, another scale invariant test of $H_0 : \beta_i = 0$ would be a likelihood ratio test, as we now show. The empirical likelihood ratio test statistic is defined as

$$\Lambda = \frac{\sup_{\gamma>0} L(0,\gamma)}{\sup_{-\infty<\beta<\infty,\gamma>0} L(\beta,\gamma)}, \tag{3.1}$$

where $L(\beta,\gamma) = \prod_{j=1}^{n_i} \frac{1}{\gamma} f(\frac{L_{ij}-\beta}{\gamma}) = \gamma^{-n_i} \prod_{j=1}^{n_i} f(\frac{L_{ij}-\beta}{\gamma})$.

When the null hypothesis is true, the likelihood function can be written as

$$\begin{aligned} L(\beta,\gamma) &= \prod_{j=1}^{n_i} \frac{1}{\gamma} f\left(\frac{L_{ij}-\beta}{\gamma}\right) \\ &= {\gamma_i}^{-n_i} (\gamma/\gamma_i)^{-n_i} \prod_{j=1}^{n_i} f\left(\frac{\epsilon_{ij}-\beta/\gamma_i}{\gamma/\gamma_i}\right) \\ &= {\gamma_i}^{-n_i} \eta^{-n_i} \prod_{j=1}^{n_i} f\left(\frac{\epsilon_{ij}-\delta}{\eta}\right), \end{aligned}$$

where $\delta = \beta/\gamma_i$ and $\eta = \gamma/\gamma_i$.

The test statistic can thus be written as

$$\Lambda = \frac{\sup_{\eta>0} L(0,\eta|\epsilon_{i1},\ldots\ \epsilon_{in_i})}{\sup_{-\infty<\delta<\infty,\eta>0} L(\delta,\eta|\epsilon_{i1},\ldots,\epsilon_{in_i})}$$

under the null hypothesis, and hence is invariant to the unknown location and scale parameters. Therefore, the distribution of the likelihood ratio statistic is completely

determined by f. Of course, we usually do not know f, but the methodology of chapter II provides an estimate of f, call it $\hat{f}$. We thus propose the use of an empirical likelihood ratio test in which f in (3.1) is replaced by $\hat{f}$. The distribution of this statistic under H_0 can be approximated using the bootstrap algorithm described in the next section.

3.3 Bootstrap Methodology for HDLSS

Given a test statistic for testing the hypothesis $H_0 : \beta_i = 0$, we need to know the sampling distribution of the test statistic under the null hypothesis. A test statistic is a function of the observations, which we assume are i.i.d. from F. It is often not easy to derive the statistic's sampling distribution even if F is known. One can approximate the sampling distribution of a given test statistic by generating many samples from F, obtaining the value of the test statistic for each of these samples, and then computing the empirical distribution of these values. However, F is usually unknown. The bootstrap method, proposed by Efron (1979) , is a method of approximating the sampling distribution of a test statistic using only information in the observed data. This is done by replacing F by $\hat{F}_n$, the empirical distribution, in the procedure just described. A summary of theoretical properties of the bootstrap is provided by Hall (1992) .

In HDLSS, the model is defined as

$$L_{ij} = \beta_i + \gamma_i \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, G,$$

where each ϵ_{ij} is i.i.d. from f. Our interest is in testing the hypotheses $H_0 : \beta_i = 0,\ i = 1, \ldots, G$, where β_i is the population mean for the ith subject. The t-statistic and a likelihood ratio statistic are considered as test statistics. As mentioned before, the distribution of each of these statistics is completely determined by f, so that we

can estimate its sampling distribution by applying a bootstrap algorithm to $\hat{f}$.

Let S be a test statistic and s an observed value of S. The steps of our bootstrap methodology for testing $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i > 0$ are as follows.

- Draw a random sample of size n_i from $\hat{f}$, and compute the statistic S. Call the value S^*.
- Repeat the previous step B times independently, yielding test statistics $S_1^*, \ldots, S_B^*$.
- The estimated p-value is $P = \frac{1}{B}\sum_{i=1}^{B} I(S_i^* \geq s)$, and H_0 is rejected at level α if $P \leq \alpha$.

We now describe how to draw samples from the kernel estimate having the form

$$\begin{aligned}
\hat{f}_h(x) &= \frac{1}{Nh}\sum_{i=1}^{G}\sum_{k=1}^{C}\sum_{j=1}^{n_i}\alpha_{ki}K\left(\frac{x-(L_{ij}-\mu_k)/\sigma_k}{h}\right) \\
&= \frac{1}{Nh}\sum_{i=1}^{G}\sum_{k=1}^{C}\alpha_{ki}\sum_{j=1}^{n_i}K\left(\frac{x-(L_{ij}-\mu_k)/\sigma_k}{h}\right) \\
&= \sum_{i=1}^{G}\sum_{k=1}^{C}\frac{n_i}{N}\alpha_{ki}\sum_{j=1}^{n_i}\frac{1}{n_i h}K\left(\frac{x-(L_{ij}-\mu_k)/\sigma_k}{h}\right) \\
&= \sum_{i=1}^{G}\sum_{k=1}^{C}\frac{n_i}{N}\alpha_{ki}\hat{f}_h(x;i,k). \qquad (3.2)
\end{aligned}$$

The equation (3.2) is a mixture of the $G \cdot C$ densities $\hat{f}_h(x;i,k),\ i = 1,\ldots,G,\ k = 1,\ldots,C$.

Therefore, one way to select bootstrap samples from $\hat{f}_h$ is as follows:

- Randomly select a subject from $(1,\ldots,G)$ with probabilities $(\frac{n_1}{N},\ldots,\frac{n_G}{N})$, and call it subject i.
- Compute $\alpha_{1i},\ldots,\alpha_{Ci}$.
- Choose a cluster with probabilities $(\alpha_{1i},\ldots,\alpha_{Ci})$, and call it k.

- The kernel estimate $\hat{f}_h(x;i,k)$ is a convolution of a $N(0,h^2)$ distribution, and the empirical distribution of $(L_{i1}-\mu_k)/\sigma_k,\ldots,(L_{in_i}-\mu_k)/\sigma_k$. To draw a sample from $\hat{f}_h(\cdot;i,k)$, we thus draw a value X^* from $N(0,h^2)$ and then randomly select one of the n_i data values, call it L_{iJ}. The value selected from $\hat{f}_h$ is then

$$X^* + \frac{L_{iJ}-\mu_k}{\sigma_k}.$$

Recall that we discussed various clustering methods in chapter II. Information obtained when the data are clustered can be used when applying the bootstrap. Each subject is assigned to one of the clusters, and bootstrap methodology for testing $H_{0i}:\beta_i=0,\ i=1,\ldots,G$, is as follows.

- Standardize each measurement by subtracting its cluster mean and dividing by its cluster standard deviation. The mean and variance of a cluster are obtained by simply computing the sample mean and variance of the data within the cluster.

- Compute the standardized data

$$e_{rs}=\frac{L_{rs}-\hat{\mu}_{k(r)}}{\hat{\sigma}_{k(r)}},\quad s=1,\ldots,n_r,\quad r=1,\ldots,G,$$

  where $k(r)$ is the cluster to which subject r belongs.

- Draw a random sample of size n_i with replacement from the N standardized values. Denote this sample $\epsilon_1^*,\ldots,\epsilon_{n_i}^*$.

- Compute the test statistic S_1^* from $\epsilon_1^*,\ldots,\epsilon_{n_i}^*$.

- Repeat previous steps for all distinct n_i, $i=1,\ldots,G$.

- Obtain p-values for all G subjects.

A third way to bootstrap is to draw samples from a unimodal density estimate, as described in section 2.2.2.1. Here we take advantage of the fact that the unimodal density estimate is piecewise constant, and use the probability integral transform to generate a value from the density.

- Let $I_1, \ldots, I_k$ be the disjoint intervals on which the density estimate is constant. Randomly select an interval, where the probability of an interval is the area of the density estimate over that interval.
- Let (x_l, x_u) be the interval. Select ϵ_1^* from $U(x_l, x_u)$.
- Repeat the previous steps until getting $(\epsilon_1^*, \ldots, \epsilon_{n_i}^*)$, and compute S_1^* from these observations.
- Repeat the previous steps B times independently, yielding test statistics $S_1^*, \ldots, S_B^*$.
- Repeat the previous steps for all distinct $n_i,\ i = 1, \ldots, G$.
- Obtain p-values of all G tests.

3.4 False Discovery Rate

Hypothesis tests on HDLSS data involve simultaneous testing of G hypotheses $H_i,\ i = 1, \ldots, G$. If we test G hypotheses with level α and all H_0's are true,

$$\begin{aligned} \textit{Experimentwise error rate} &= P(\text{at least 1 hypothesis is falsely rejected}) \\ &= 1 - (1-\alpha)^G = \alpha_G. \end{aligned}$$

This last probability is much bigger than α when G is even moderately large. Controlling experimentwise error rate to be small, say, 0.05, leads to very conservative tests. Instead, one may control FDR as defined in (3.3).

When G hypotheses are tested, all possible situations for simultaneous tests can

be summarized as in Table 1 (Benjamini and Hochberg, 1995). It is assumed that G, the number of hypotheses, is known, G_0, the number of true null hypotheses, and G_1,the number of non-true null hypotheses, are unknown parameters. The number of rejected, R, is an observable random variable and A,B,C and D as defined in Table 1 are unobservable random variables (Dudoit et al., 2003).

The false discovery rate (FDR) of Benjamini and Hochberg (1995) is the expected proportion of Type I errors among the rejected hypotheses, that is,

$$FDR = E(Q), \tag{3.3}$$

where, by definition, $Q = B/R$ if $R > 0$ and 0 if $R = 0$.

We define *power* in the simultaneous testing context to be

$$Power = E(D/G_1).$$

The steps of the Benjamini and Hochberg (1995) method of controlling FDR are as follows (Benjamini and Hochberg, 1995).

- Consider hypotheses $H_1, H_2, \ldots, H_G$ and corresponding p-values $P_1, P_2, \ldots, P_G$.
- Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(G)}$ be the ordered p-values and let $H_{(i)}$ be the null hypothesis corresponding to $P_{(i)}$.
- Let k be the largest i that satisfies $P_{(i)} \leq \frac{i}{G}\alpha$.
- Reject all $H_{(i)},\ i = 1, 2, \ldots, k$.

Table 1: Possible Decisions in Simultaneous Testing of G Hypotheses

	not rejected	rejected	total
Number of null hypotheses	A	B	G_0
Number of non-true null hypotheses	C	D	G_1
hypotheses	G-R	R	G

It is shown by Benjamini and Hochberg (1995) that when $P_1, \ldots, P_G$ are independent, the above procedure controls the FDR to be α.

CHAPTER IV

SIMULATIONS AND DATA ANALYSIS

In previous chapters we proposed methods for performing hypothesis tests on each of a large number of small data sets. In this chapter we conduct simulation studies and analyze a real data set to investigate the proposed methods.

4.1 Simulation Studies

We simulate data sets from various mixture distributions and each component density has one mode, mean 0 and variance 1. The mixture of Laplace distributions, the mixture of Gamma distributions and the mixture of t-distributions with 3 degrees of freedom are considered. The true number of clusters is taken to be 3.

The Laplace mixture has the form

$$\sum_{k=1}^{3} \frac{w_k}{\sigma_k} \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}\left|\frac{x-\mu_k}{\sigma_k}\right|\right),$$

the Gamma mixture is

$$\sum_{k=1}^{3} \frac{w_k}{\sigma_k} 2\left(\frac{x-\mu_k}{\sigma_k}+\sqrt{2}\right) \exp\left(-\sqrt{2}\left(\frac{x-\mu_k}{\sigma_k}\right)\right) I\left(\frac{x-\mu_k}{\sigma_k} > -\sqrt{2}\right),$$

in which $I(A) = 1$ if A is true and $I(A) = 0$ otherwise, and the t_3 mixture has the form

$$\sum_{k=1}^{3} \frac{w_k}{\sigma_k} \frac{2}{\pi}\left(\frac{1}{1+(\frac{x-\mu_k}{\sigma_k})^2}\right),$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ and $\boldsymbol{w} = (w_1, w_2, w_3)$ are means, standard deviations and proportions, respectively. The three cluster means are $\boldsymbol{\mu} = (-2, 2, 0)$, and the cluster standard deviations are $\boldsymbol{\sigma} = (1, 1, 1)$. Two sets of cluster proportions

are considered: $\boldsymbol{w} = (0.3, 0.3, 0.4)$ and $\boldsymbol{w} = (0.15, 0.1, 0.75)$. We take G (the number of subjects) to be 800, and each subject has 3 observations.

The steps of the simulation are as follows:

- Generate 800 data sets (with 3 repetitions each) from a given mixture distribution. Observations from the same data set (or subject) are generated from the same cluster.
- Calculate the mean of sample variances over all subjects (MSV).
- Apply K-means clustering to the subject means to get initial parameter estimates and initial clusterings. This is done for each k, $k = 1, \ldots, C_{max}$.
- Calculate ($WMCV$) at each number of clusters by using the variance estimates from K-means clustering. Also, compute BIC at each number of clusters using a method to be described below. Choose the number $\hat{C}$ of clusters that has the minimum $|WMCV(C) - MSV|$ or maximizes BIC, and consider this as the optimal number of clusters. Obtain an estimate of f using either kernel density estimation or unimodal density estimation. This estimate is computed with number of clusters equal to $\hat{C}$. Initial estimates for the kernel density estimate are those obtained from K-means clustering (with $\hat{C}$ clusters).
- Obtain the bootstrap distribution of the t-statistic using the algorithm described in Section 3.3. The number of bootstrap samples is 100,000.
- Compute approximate p-values for each of the 800 subjects.
- Use the FDR method (Benjamini and Hochberg, 1995) to determine which null hypotheses should be rejected.

The procedure just described was repeated 50 times for each of the 6 mixture distributions considered. Table 2 summarizes the percentage of correct C values that is chosen by WMCV and BIC for each of the 6 mixture cases. Obviously, the WMCV method worked much better than BIC. Six plots of $WMCV$ and MSV versus # of clusters are shown in Figure 2.

In our simulation study, the second iteration, $\hat{f}_2$, of the procedure described in section 2.2.2.2 is used for $\hat{f}$. Tables 3 - 5 show the average of 50 estimated percentiles from the bootstrap distribution. The quantities $T_{0.025}$ (Bootstrap$_{\mathrm{f}}$) and $T_{0.025}$ (Bootstrap$_{\hat{\mathrm{f}}}$) denote $2.5th$ percentiles of the sampling distribution of a t-statistic as approximated by repeated sampling from f and $\hat{f}$, respectively. Parameters for Case 1 are $\boldsymbol{\mu} = (-2, 2, 0)$, $\boldsymbol{\sigma} = (1, 1, 1)$ and $\boldsymbol{w} = (0.3, 0.3, 0.4)$, and parameters for Case 2 are the same as those of Case 1 except $\boldsymbol{w} = (0.15, 0.1, 0.75)$.

Table 2: Percentage of Correctly Chosen C

Mixture, Case	WMCV	BIC
Laplace, Case 1 $\boldsymbol{w} = (0.3, 0.3, 0.4)$	100%	74%
Laplace, Case 2 $\boldsymbol{w} = (0.15, 0.1, 0.75)$	100%	52%
Gamma, Case 1 $\boldsymbol{w} = (0.3, 0.3, 0.4)$	100%	8%
Gamma, Case 2 $\boldsymbol{w} = (0.15, 0.1, 0.75)$	100%	12%
T_3, Case 1 $\boldsymbol{w} = (0.3, 0.3, 0.4)$	94%	84%
T_3, Case 2 $\boldsymbol{w} = (0.15, 0.1, 0.75)$	94%	10%

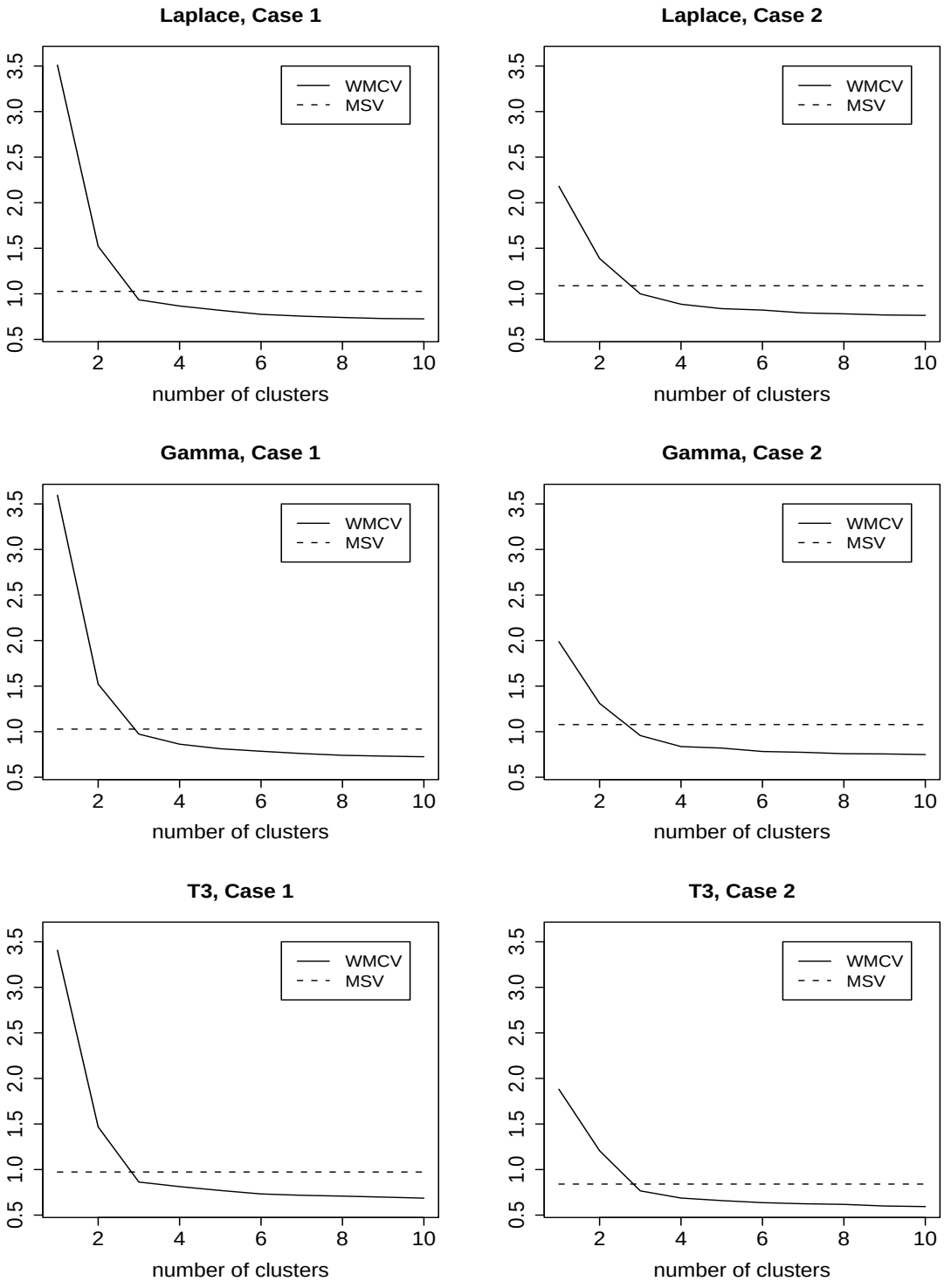

Figure 2: Plot of WMCV vs. Number of Clusters

Table 3: Estimated Percentiles of Sampling Distribution for Laplace Mixture

	Parameters	$T_{0.025}, T_{0.975}$ (Bootstrap$_{\mathrm{f}}$)	$T_{0.025}$, $T_{0.975}$ (Bootstrap$_{\hat{\mathrm{f}}}$)
Case 1	$\boldsymbol{w} = (0.3, 0.3, 0.4)$	-3.4798, 3.4739	-3.4963, 3.4739
Case 2	$\boldsymbol{w} = (0.15, 0.1, 0.75)$	-3.4712, 3.4750	-3.4752, 3.6131

Table 4: Estimated Percentiles of Sampling Distribution for Gamma Mixture

	Parameters	$T_{0.025}, T_{0.975}$ (Bootstrap$_{\mathrm{f}}$)	$T_{0.025}$, $T_{0.975}$ (Bootstrap$_{\hat{\mathrm{f}}}$)
Case 1	$\boldsymbol{w} = (0.3, 0.3, 0.4)$	-8.2173, 2.6477	-7.4076, 2.7905
Case 2	$\boldsymbol{w} = (0.15, 0.1, 0.75)$	-8.2030, 2.6376	-7.2867, 2.8353

Table 5: Estimated Percentiles of Sampling Distribution for T_3 Mixture

	Parameters	$T_{0.025}, T_{0.975}$ (Bootstrap$_{\mathrm{f}}$)	$T_{0.025}$, $T_{0.975}$ (Bootstrap$_{\hat{\mathrm{f}}}$)
Case 1	$\boldsymbol{w} = (0.3, 0.3, 0.4)$	-3.7553, 3.7382	-3.7073, 3.7540
Case 2	$\boldsymbol{w} = (0.15, 0.1, 0.75)$	-3.7437, 3.7307	-3.7832, 3.6835

For a given C, BIC is defined as follows:

$$BIC(C) = 2\log \hat{L}_C - (3C-1)\log N,$$

where N is the total number of observations, and

$$\hat{L}_C = \prod_{i=1}^{G}\prod_{j=1}^{n_i} \hat{f}_h(L_{ij}).$$

The density estimate $\hat{f}_h$ is defined by

$$\hat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{G}\sum_{j=1}^{n_j} K\left(\frac{x-e_{ij}}{h}\right),$$

where $e_{ij},\ j=1,\ldots,n_i,\ i=1,\ldots,G$, are as defined in expression (2.5). All parameters are estimated by K-means clustering. Now one may calculate BIC for each C, and the number of clusters which maximizes BIC is chosen as an optimal number of clusters.

P-values corresponding to the 800 hypotheses are calculated, and the empirical cdf of these p-values for case 1 of the Laplace mixture is given in Figure 3. Similar plots for the other cases are shown in Appendix E. As expected, the empirical cdf of p-values under the null hypothesis (cluster 3) is approximately uniform and different from those under the alternative hypotheses (clusters 1 and 2). It is obviously more difficult to reject the hypothesis when a subject comes from the null hypothesis.

The Benjamini-Hochberg method of controlling FDR is applied, with the nominal FDR 0.05. For each set of 800 data sets, the Q is calculated by

$$Q = \begin{cases} \frac{\#\text{ of rejected true null hypotheses}}{\#\text{ of rejected}}, & \text{if } \#\text{ of rejected} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

and the *power* is defined as

$$Power = \frac{\#\text{ of rejected non} - \text{true null hypotheses}}{\#\text{ of non true null hypotheses}}.$$

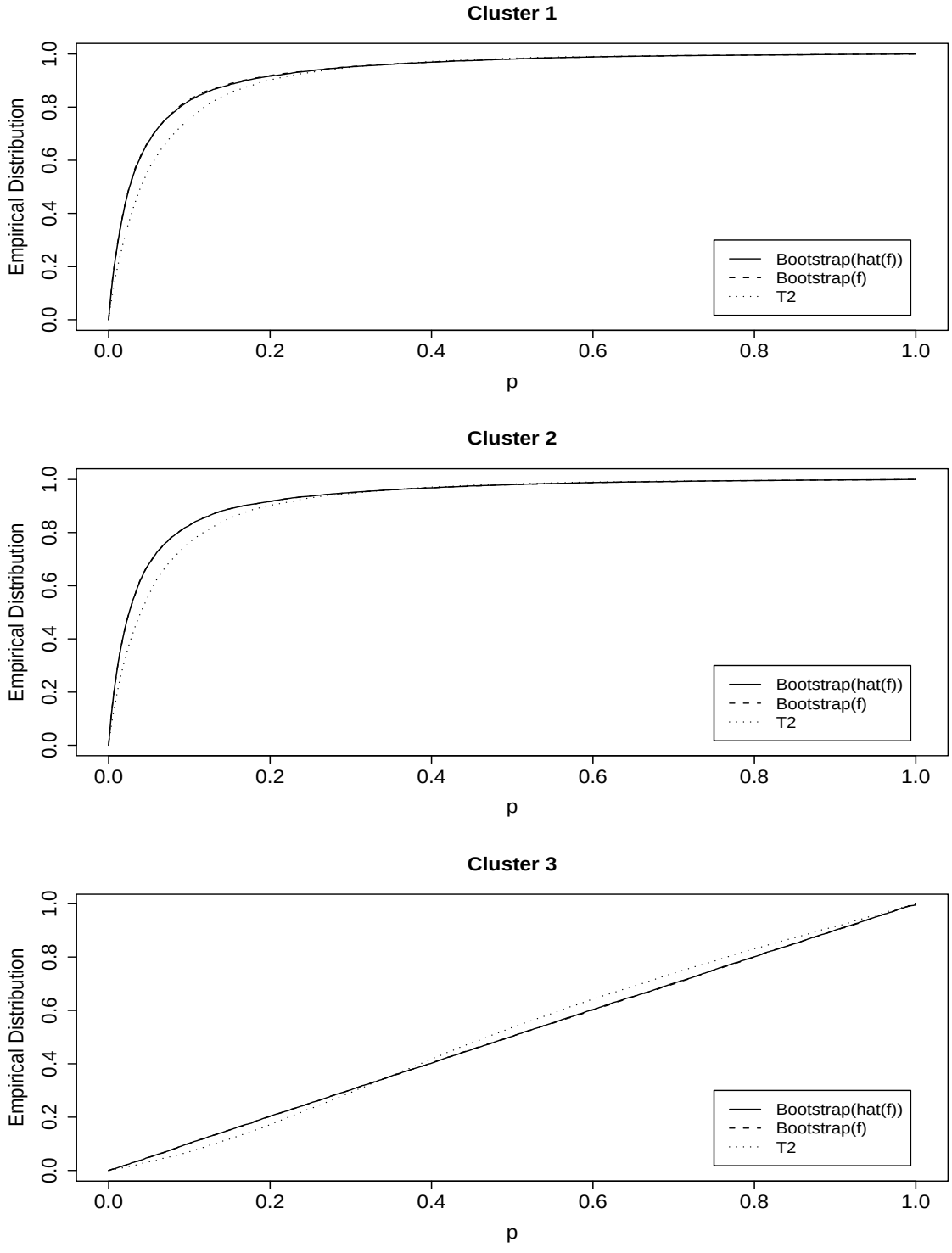

Figure 3: Empirical Distribution of p-values for Case 1 of Laplace Mixture

Tables 6 - 8 summarize the average of 50 Qs and *powers* for each mixture distribution. Three different sets of percentiles for the t-statistic are provided by $\text{Bootstrap}_{\text{f}}$, $\text{Bootstrap}_{\hat{\text{f}}}$ and T_2. $\text{Bootstrap}_{\text{f}}$ and $\text{Bootstrap}_{\hat{\text{f}}}$ denote the sampling distribution of the t-statistic approximated by resampling from f and $\hat{f}$, respectively, while T_2 uses the t-distribution with 2 degrees of freedom as the distribution of the t-statistic, since each small data set has 3 repetitions.

In most cases, the p-values calculated by $\text{Bootstrap}_{\hat{\text{f}}}$ are close to the p-values based on knowledge of the true distribution ($\text{Bootstrap}_{\text{f}}$), and the Q is less than 0.05, except T_2 for case 1 of Gamma mixture . As we see in Tables 6 - 8, $\text{Bootstrap}_{\hat{\text{f}}}$ works reasonably well in most cases since it has powers similar to those of $\text{Bootstrap}_{\text{f}}$ and better powers than T_2. For case 2 of Gamma mixture, the power obtained by T_2 is slightly bigger than that of $\text{Bootstrap}_{\hat{\text{f}}}$ and $\text{Bootstrap}_{\text{f}}$, but the Q of T_2 is much bigger than that of $\text{Bootstrap}_{\hat{\text{f}}}$ and $\text{Bootstrap}_{\text{f}}$.

4.2 Real Data Analysis

A real microarray data set provided by Dr. Kenneth Ramos and Dr. Charlie D. Johnson, both formerly of Texas A&M University is analyzed. The data are obtained from a cDNA microarray experiment, and there are 813 genes of interest with gene expressions under treatment and control conditions. Most genes have 3 repetitions for both conditions. An appropriate background subtraction and normalization has been done to ensure that any significant difference between treatment and control is due to a treatment, as opposed to a dye effect.

Let L_{ij} be the jth measurement for the ith gene, as defined by

$$L_{ij} = \ln(T_{ij}/C_{ij}), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, 813,$$

where T_{ij} and C_{ij} are the jth expression for the ith gene under the treatment and

Table 6: Q and Power of Laplace Mixture

	Q			Power		
	$\text{Bootstrap}_{\text{f}}$	$\text{Bootstrap}_{\hat{\text{f}}}$	T_2	$\text{Bootstrap}_{\text{f}}$	$\text{Bootstrap}_{\hat{\text{f}}}$	T_2
Case 1	0.0202	0.0178	0.0080	0.1495	0.1462	0.0150
Case 2	0.0495	0.0211	0.0150	0.0113	0.0076	0.0025

Table 7: Q and Power of Gamma Mixture

	Q			Power		
	$\text{Bootstrap}_{\text{f}}$	$\text{Bootstrap}_{\hat{\text{f}}}$	T_2	$\text{Bootstrap}_{\text{f}}$	$\text{Bootstrap}_{\hat{\text{f}}}$	T_2
Case 1	0.0297	0.0308	0.0546	0.0117	0.0089	0.0079
Case 2	0.0161	0.0225	0.0473	0.0084	0.0058	0.0093

Table 8: Q and Power of T_3 Mixture

	Q			Power		
	$\text{Bootstrap}_{\text{f}}$	$\text{Bootstrap}_{\hat{\text{f}}}$	T_2	$\text{Bootstrap}_{\text{f}}$	$\text{Bootstrap}_{\hat{\text{f}}}$	T_2
Case 1	0.0227	0.0246	0.0075	0.1816	0.1840	0.0214
Case 2	0.0245	0.0461	0.0240	0.0071	0.0079	0.0029

control conditions, respectively. We assume that

$$L_{ij} = \beta_i + \gamma_i \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, 813,$$

where ϵ_{ij}'s are i.i.d. as f, and we estimate f by the method presented in section 2.2.2.2.

We want to compare gene expressions under treatment to control for each gene, and define the hypotheses $H_{0i} : \beta_i = 0$ vs. $H_{1i} : \beta_i \neq 0$ $(i = 1, \ldots, 813)$.

Figure 4 is a kernel density estimate of $L_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, G$, and it indicates that using a mixture distribution is desirable since it has several modes. The form of the kernel density estimate is

$$\hat{f}(x) = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{1}{Nh} K\left(\frac{x - L_{ij}}{h}\right),$$

where $N = \sum_{i=1}^{G} \sum_{j=1}^{n_i}$, K is the Gaussian kernel, $h = 1.06\hat{\sigma}N^{-1/5}$ and $\hat{\sigma}$ is the sample standard deviation of the data. Figure 5 is a scatter plot of the log(Variance) versus the mean for all 813 genes.

We assume that the L_{ij}'s are a random sample from a mixture distribution of the form

$$f_{\text{mixture}}(x) = \sum_{k=1}^{C} \frac{1}{w_k} f\left(\frac{x - \mu_k}{\sigma_k}\right)$$

where f is a density with mean 0 and standard deviation 1.

It is important to know the correct number of components, C. First we calculate the mean of the sample variances of all 813 genes (MSV) to apply the method of estimating C. The MSV is 0.0588.

K-means clustering is applied to the 813 means. Estimates $\hat{\sigma_1}^2, \ldots, \hat{\sigma_k}^2$ of the cluster variances are thereby obtained for each number k of clusters from 1 to 20. The $WMCV$ at each number of clusters is compared with the MSV, and the number of clusters having the $WMCV$ closest to the MSV is taken to be the estimate

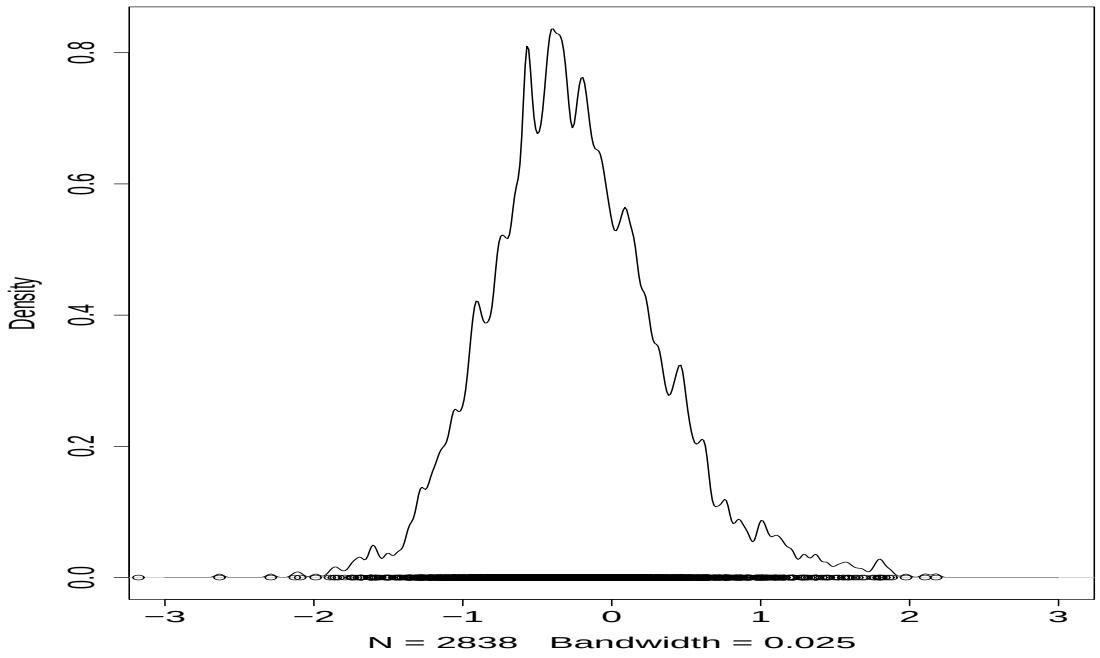

Figure 4: Kernel Density Estimation of L_{ij}'s

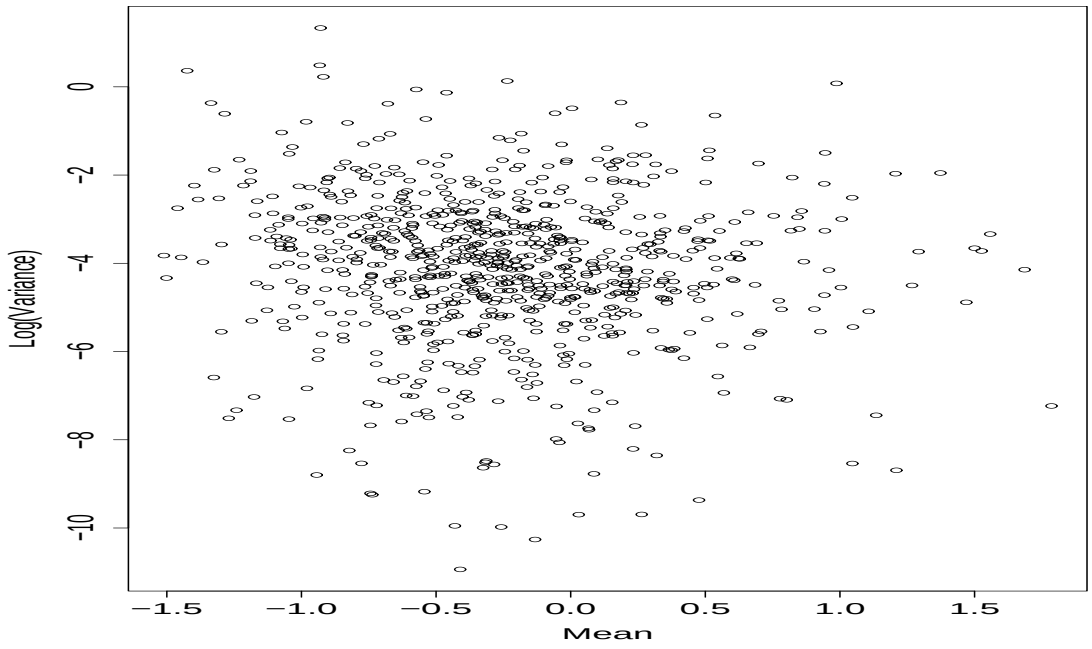

Figure 5: Plot of log(Variance) vs. Mean of Each Gene

of C. We repeat the procedure of finding C by using K-means clustering 100 times, since K-means starts with arbitrary initial clustering and will end up with different clustering results sometimes.

The estimate of C turned out to be 14 by WMCV, since 14, 13 and 15 were chosen by WMCV 72%, 23% and 5%, respectively, out of 100 times. Figure 6 shows a plot of $WMCV$ versus # of clusters for one of the K-means clustering results, and the dotted and the solid lines represent MSV and $WMCV$, respectively. The optimal K-means clustering denotes that one of the K-means clustering results corresponds to the minimum of the $|WMCV(C) - MSV|$ and $C = 14$. BIC is also applied and it ends up with $C = 1$ all the time. The procedure of BIC is the same as that described in section 4.1. We took C to be 14 on the basis of the WMCV results.

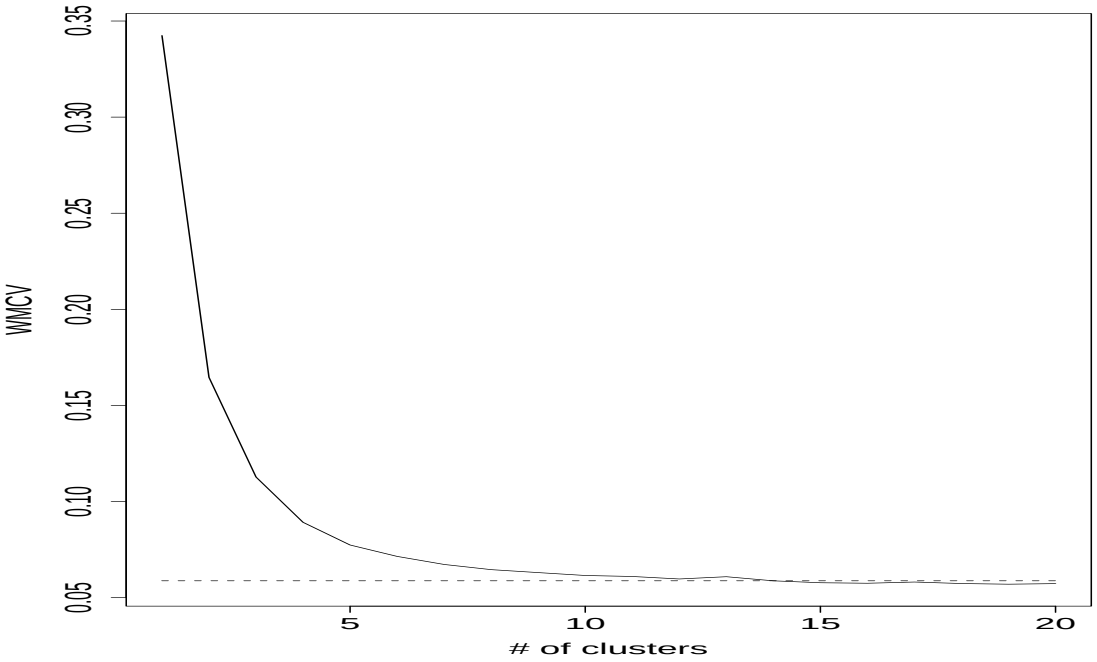

Figure 6: Plot of WMCV vs. Number of Clusters

The density f is estimated by a kernel estimator, and two initial clustering methods, K-means clustering and the Gaussian mixture model are applied. We iterate $\hat{f}_r$ by updating parameter estimates, where $\hat{f}_r$ is the estimate of f at the rth iteration.

The steps of estimating f based on K-means clustering result is as follows:

- Get $\hat{f}_1$ by using optimal K-means clustering and $C = 14$. The data are standardized as described on p. 37, and $\hat{f}_0$ is a standard kernel estimate computed from these data or obtained by $\hat{f}_{h,0}$ on p. 21.
- Iterate $\hat{f}_r$, $(r \leq 12)$ as described in section 2.2.2.2 until it converges. The resulting iterates are shown in Figure 7, and the solid and the dotted lines are $\hat{f}_r$ and ϕ, respectively.
- Consider either $\hat{f}_{12}$ or $\hat{f}_{converged}$ as $\hat{f}$.

The steps of estimating f based on an initial Gaussian mixture model are as follows:

- Apply agglomerative clustering with a normal assumption to get an initial estimate of the Gaussian mixture.
- Let $\hat{f}_1$ be the fitted Gaussian mixture model with $C = 14$.
- Iterate $\hat{f}_r$, $(r \leq 12)$ until it converges. The iterates are shown in Figure 8, and the solid and the dotted lines are $\hat{f}_r$ and ϕ, respectively.
- Consider either $\hat{f}_{12}$ or $\hat{f}_{converged}$ as $\hat{f}$.

Finally we obtain 100,000 bootstrap samples by drawing samples from $\hat{f}$. For each set of bootstrap samples, a t-statistic is calculated and the sampling distribution of the statistic is approximated based on these values.

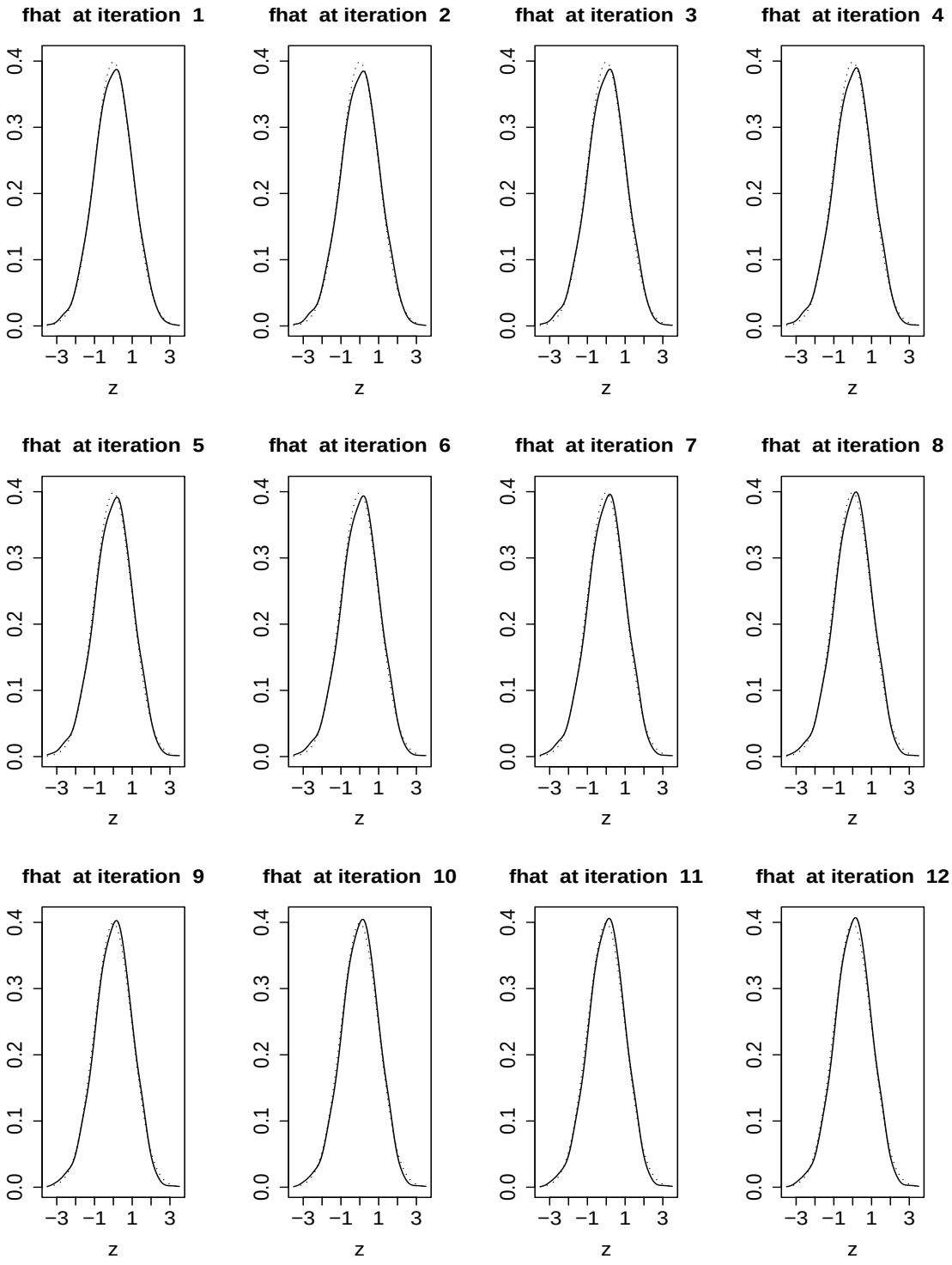

Figure 7: $\hat{f}_r$ at Each Iteration with K-means Initial

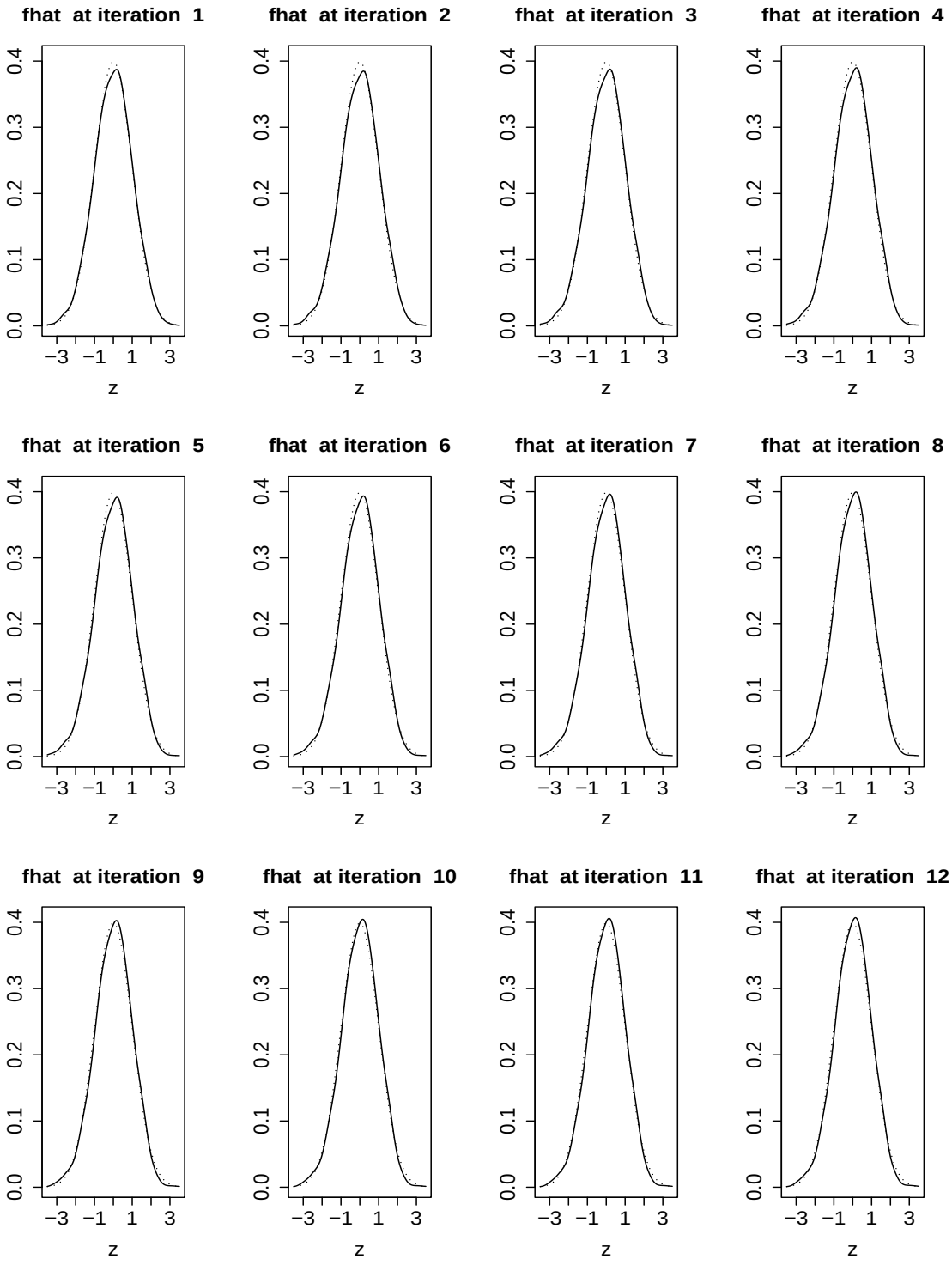

Figure 8: $\hat{f}_r$ at Each Iteration with Gaussian Mixture Initial

Recall that the hypotheses $H_{0i} : \beta_i = 0$ vs. $H_{1i} : \beta_i \neq 0$ need to be tested. For 813 subjects, we obtain 813 observed t-values and the corresponding p-values, as approximated by the bootstrap. By using the procedure that controls FDR, we decide which gene is significantly expressed under the treatment as compared to the control condition. FDR is controlled at $\alpha = 0.05$.

Tables 9 - 10 summarize the results of bootstrapping. Three different sampling distributions of the t-statistic are compared: $\text{Bootstrap}_{\text{K}}$ and $\text{Bootstrap}_{\text{G}}$ denote the bootstrap sampling distribution of the t-statistic when $\hat{f}$ is based on K-means clustering and Gaussian mixture modeling, respectively, and T_{df} uses the t-distribution with degrees of freedom equal to (# of repetitions for each subject) $- 1$. As shown from the percentage of rejections in Table 10, tests based on the bootstrap appear to be more powerful than tests based on the t-distribution.

About 49.2%, 47.9% and 42.8% of the hypotheses are rejected based on $\text{Bootstrap}_{\text{K}}$, $\text{Bootstrap}_{\text{G}}$ and T_{df}, respectively.

Table 9: Estimated Percentiles

Sample size	$\text{Bootstrap}_{\text{K}}$	$\text{Bootstrap}_{\text{G}}$	T_{df}
3	-3.760631, 4.580793	-4.009903, 4.656244	-4.302653, 4.302653
6	-2.353795, 2.716064	-2.402902, 2.740441	-2.570582, 2.570582
9	-2.128096, 2.443206	-2.182742, 2.396822	-2.306004, 2.306004
12	-2.054145, 2.312548	-2.087544, 2.312441	-2.200985, 2.200985

Table 10: Number of Rejected the Null Hypothesis out of 813 Genes

	$\text{Bootstrap}_{\text{K}}$	$\text{Bootstrap}_{\text{G}}$	T_{df}
Number of rejected	400	389	348
Percentage of rejected	49.20%	47.85%	42.80%

CHAPTER V

CONCLUSIONS AND FUTURE STUDIES

5.1 Summary

This dissertation has been concerned with hypothesis tests on small data sets which come from HDLSS data. Microarray data are one example of such data.
We proposed a general statistical model to express the jth measurement of the ith subject from HDLSS data.

One of many interesting problems in microarray analysis is detecting genes with significant expression under treatment as compared to control conditions. Quantities $\log(\frac{T}{C})$ are obtained to address this problem and the statistical model as in (2.1) is applied to these quantities, where T and C are measurements from treatment and control, respectively. By testing an appropriate hypothesis for each gene, we may decide which genes are significantly affected by the treatment. The statistical model and testing procedure are not limited to comparative experiments. They can be used for HDLSS data in general.

In this dissertation, we proposed the WMCV method for choosing the number of clusters, and simulation studies show that it worked quite well as it selected the true number of clusters in 294 of 300 data sets. Another important methodology is a modified kernel estimate of f proposed in section 2.2.2.2. This estimator promises to be consistent under general conditions, as evidenced by approximations of various mixtures; see Appendices A-D. We also proposed a new way of bootstrapping using the modified kernel estimate of section 3.3. Simulation studies showed that this bootstrap provides results comparable to those obtained when using knowledge of the true underlying distribution. In addition, the bootstrap yielded better power

than tests based on the assumption that the true underlying distribution is Gaussian.

We used the method of controlling FDR (Benjamini and Hochberg, 1995) at $\alpha = 0.05$, and most bootstrap simulation results were such that the empirical false discovery rate was less than 0.05.

We analyzed a set of cDNA microarray data from a comparative experiment. By using WMCV, we chose 14 as an optimal number of clusters, and the kernel density estimator at 14 clusters was obtained. The percentages of genes determined to be significantly expressed under treatment by $\text{Bootstrap}_{\text{K}}$, $\text{Bootstrap}_{\text{G}}$ and T_{df} were 49.20%, 47.85% and 42.80%, respectively.

5.2 Future Studies

In this dissertation we considered a mixture model by assuming each component density is the same up to location and scale. One can extend this by allowing a more general family of distributions. One possibility is to allow the components to be arbitrary unimodal densities.

The consistency of $\hat{f}_r$ should also be studied further. We provided evidence that $E(\hat{f}_r) = f_r$ converges to f as $r \to \infty$, but did not show that $Var(\hat{f}_r) \to 0$ as $Nh \to \infty$. Therefore, the behavior of $Var(\hat{f}_r)$ at each iteration is also an interesting problem to consider. We used the true parameter values in defining the $\hat{f}_r$ whose expectation was investigated. One may add the uncertainty of estimating parameters when investigating consistency of a kernel estimator.

In our density estimation method, initial clustering is important, so by considering other clustering methods than those that we used, one may get better performance of a density estimator. One possibility is to use a random search to find a grouping, or clustering, that maximizes the likelihood for our mixture model when it is assumed only that the density f is an arbitrary unimodal density.

The empirical likelihood ratio statistic proposed in chapter III is also worthy of more study. If f can be well-estimated, this likelihood ratio test has the promise of much better power than a t-test, at least for many densities f.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing." *Journal of Royal Statistical Society, Series B*, 57, 289–300.

Conzone, S. D. and Pantanot, C. G. (2004). "Glass Slides to DNA Microarrays." *Matrialstoday*, 20–26.

Cui, X. and Churchill, G. (2003). "Statistical Tests for Differential Expression in cDNA Microarray Experiments." *Genome Biology*, 4, 210.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). "Multiple Hypothesis Testing in Microarray Experiments." *Statistical Science*, 18, 71–103.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jacknife." *The Annals of Statistics*, 7, 1–26.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). "Empirical Bayes Analysis of a Microarray Experiment." *Journal of the American Statistical Association*, 96, 1151–1160.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). "Cluster Analysis and Display of Genome-wide Expression Patterns." *Proceedings of the National Academy Sciences*, 95, 14863–14868.

Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association*, 97, 611–629.

Hall, P. G. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hall, P. G., Marron, J. S., and Neeman, A. (2003). "Geometric Representation of High Dimension Low Sample Size Data." Available at http://www.stat.unc.edu/postscript/papers/marron/GeomRepn/hm5.pdf.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

Hartigan, J. A. and Wong, M. A. (1979). "A K-means Clustering Algorithm." *Applied Statistics*, 28, 100–108.

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002). "Bayesian Models for Gene Expression with DNA Microarray Data." *Journal of the American Statistical Association*, 97, 88–99.

Ishwaran, H. and Rao, J. S. (2003). "Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection." *Journal of the American Statistical Association*, 98, 438–455.

Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., and Churchill, G. A. (2002). "Statistical Analysis of a Gene Expression Microarray Experiment with Replication." *Statistical Sinica*, 12, 203–217.

Klebanov, L., Gordon, A., Xiao, Y., Land, H., and Yakovlev, A. (2004). "A New Test Statistic for Testing Two-Sample Hypotheses in Microarray Data Analysis." Tech. Rochester Technical Report.

McLachlan, G. J., Bean, R. W., and Peel, E. (2002). "A Mixture Model-based Approach to the Clustering of Microarrayexpression Data." *Bioinformatics*, 18, 413–422.

Meyer, M. C. (2001). "An Alternative Unimodal Density Estimator with a Consistent Estimate of the Mode." *Statistica Sinica*, 11, 1159–1174.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Storey, J. (2003). "The Positive False Discovery Rate: A Bayesian interpretation and the q-Value." *The Annals of Statistics*, 31, 2013–2035.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). "Systematic Determination of Genetic Network Architecture." *Nature Genetics*, 22, 281–285.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response." *Proceedings of the National Academy Sciences*, 98, 5116–5121.

Van del Laan, M. and Bryan, J. (2001). "Gene Expression Analysis With the Parametric Bootstrap." *Biostatistics*, 2, 445–461.

Wegman, E. J. (1969). "A Note on Estimating a Unimodal Density." *The Annals of Mathematical Statistics*, 40, 1661–1667.

— (1970). "Maximum Likelihood Estimation of a Unimodal Density Function." *The Annals of Mathematical Statistics*, 41, 457–471.

SUPPLEMENTAL SOURCES

Bhattacharya, C. G. (1967). "A Simple Method of Resolution of a Distribution into Gaussian Components." *Biometrics*, 23, 115–135.

Cavanaugh, J. E. and Neath, A. A. (1999). "Generalizing the Derivation of the Schwarz Information Criterion." *Communications in Statistics*, 28, 49–66.

Cheng, R. C. H. and Liu, W. B. (2001). "The Consistency of Estimators in Finite Mixture Models." *The Scandinavian Journal of Statistics*, 28, 603–616.

Day, N. E. (1969). "Estimating the Components of a Mixture of Normal Distributions." *Biometrika*, 56, 463–474.

Didier, G., Brezellec, P., Remy, E., and A.Henaut (2002). "GeneANOVA - Gene Expression Analysis of Variance." *Bioinformatics*, 18, 490–491.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association*, 97, 77–87.

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000). "Microarrays and Their Use in a Comparative Experiment." Tech. Stanford Technical Report.

Hall, P. G. (1990). "Pseudo-Likelihood Theory for Empirical Likelihood." *The Annals of Statistics*, 18, 121–140.

Huang, X. and Pan, W. (2003). "Linear Regression and Two-class Classification with Gene Expression Data." *Bioinformatics*, 19, 2072–2078.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001). "Consistent Estimation of Mixture Complexity." *The Annals of Statistics*, 29, 1281–1296.

Kerr, M. K. and Churchill, G. A. (2001). "Experimental Design for Gene Expression Microarrays." *Biostatistics*, 2, 183–201.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). "Analysis of Variance for Gene Expression Microarray Data." *Journal of Computational Biology*, 7, 819–837.

Marriott, F. H. C. (1971). "Practical Problems in a Method of Cluster Analysis." *Biometrics*, 27, 501–514.

Redner, R. (1981). "Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions." *The Annals of Statistics*, 9, 225–228.

Robertson, T. (1967). "On Estimating a Density Which is Measurable with Respect to a σ-Lattice." *The Annals of Mathematical Statistics*, 38, 482–493.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, 6, 461–464.

Shao, J. (1999). *Mathematical Statistics*. New York: Springer.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Silverman, B. W. and Young, G. A. (1987). "The Bootstrap : To Smooth or Not to Smooth?" *Biometrika*, 74, 469–479.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Whittle, P. (1958). "On the Smoothing of Probability Density Functions." *Journal of Royal Statistical Society, Series B*, 20, 334–343.

Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002). "Comparison of Methods for Image Analysis on cDNA Microarray Data." *Journal of Computational and Graphical Statistics*, 11, 108–136.

APPENDIX A

PLOTS OF LAPLACE, GAMMA AND T_3 MIXTURES

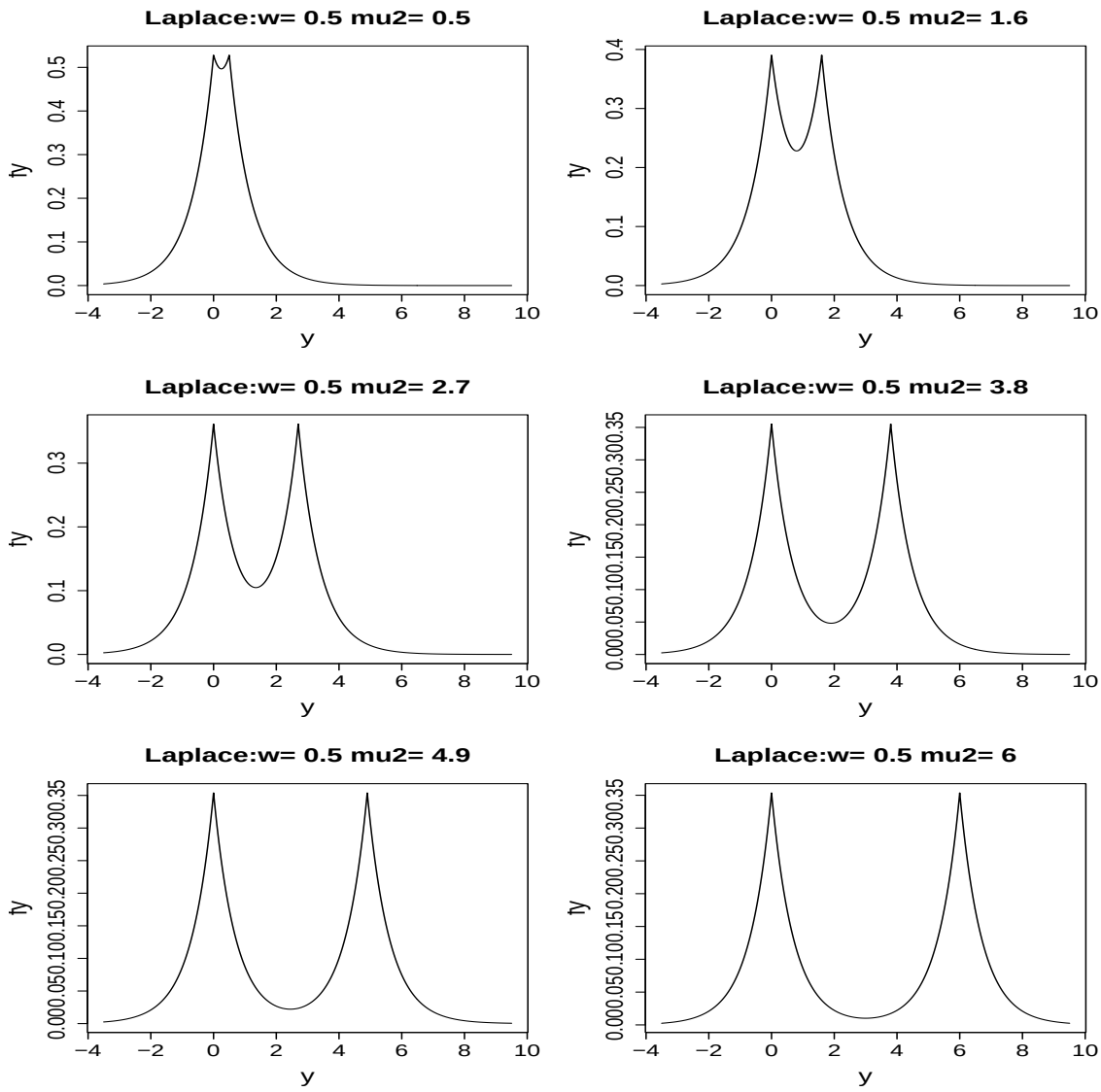

Figure 9: Mixture of Laplace Densities : $w_1 = 0.5$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

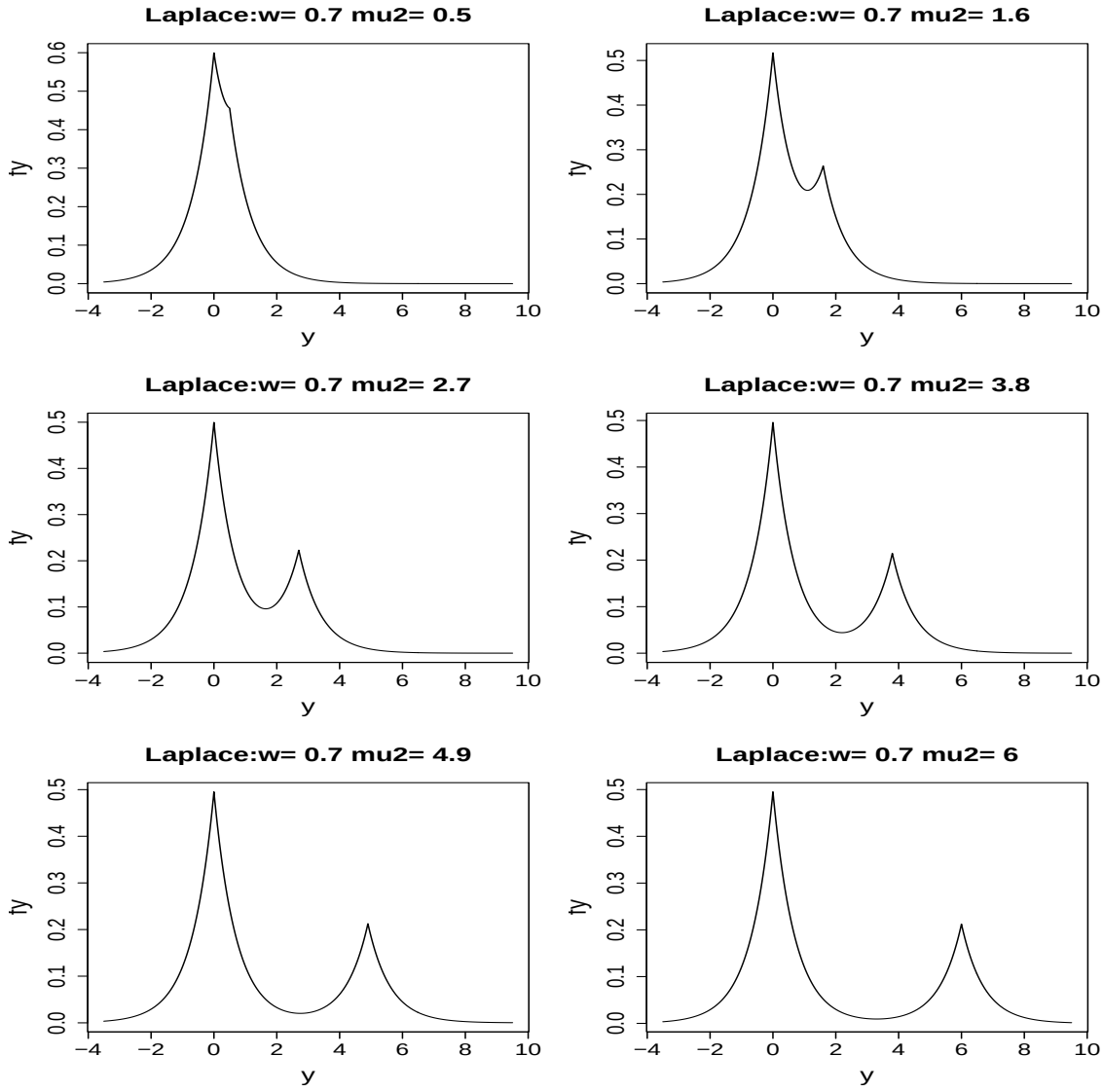

Figure 10: Mixture of Laplace Densities : $w_1 = 0.7$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

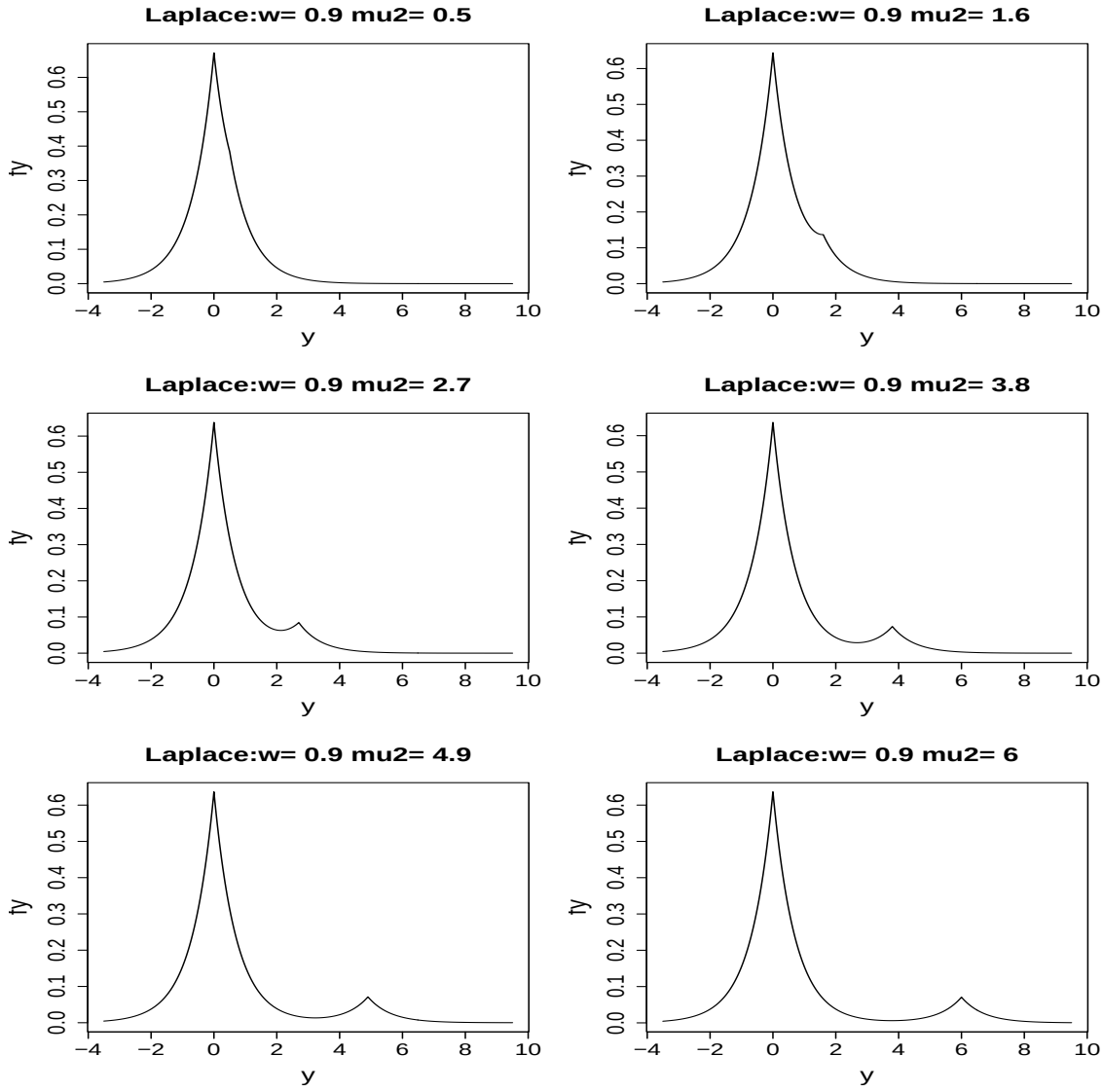

Figure 11: Mixture of Laplace Densities : $w_1 = 0.9$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

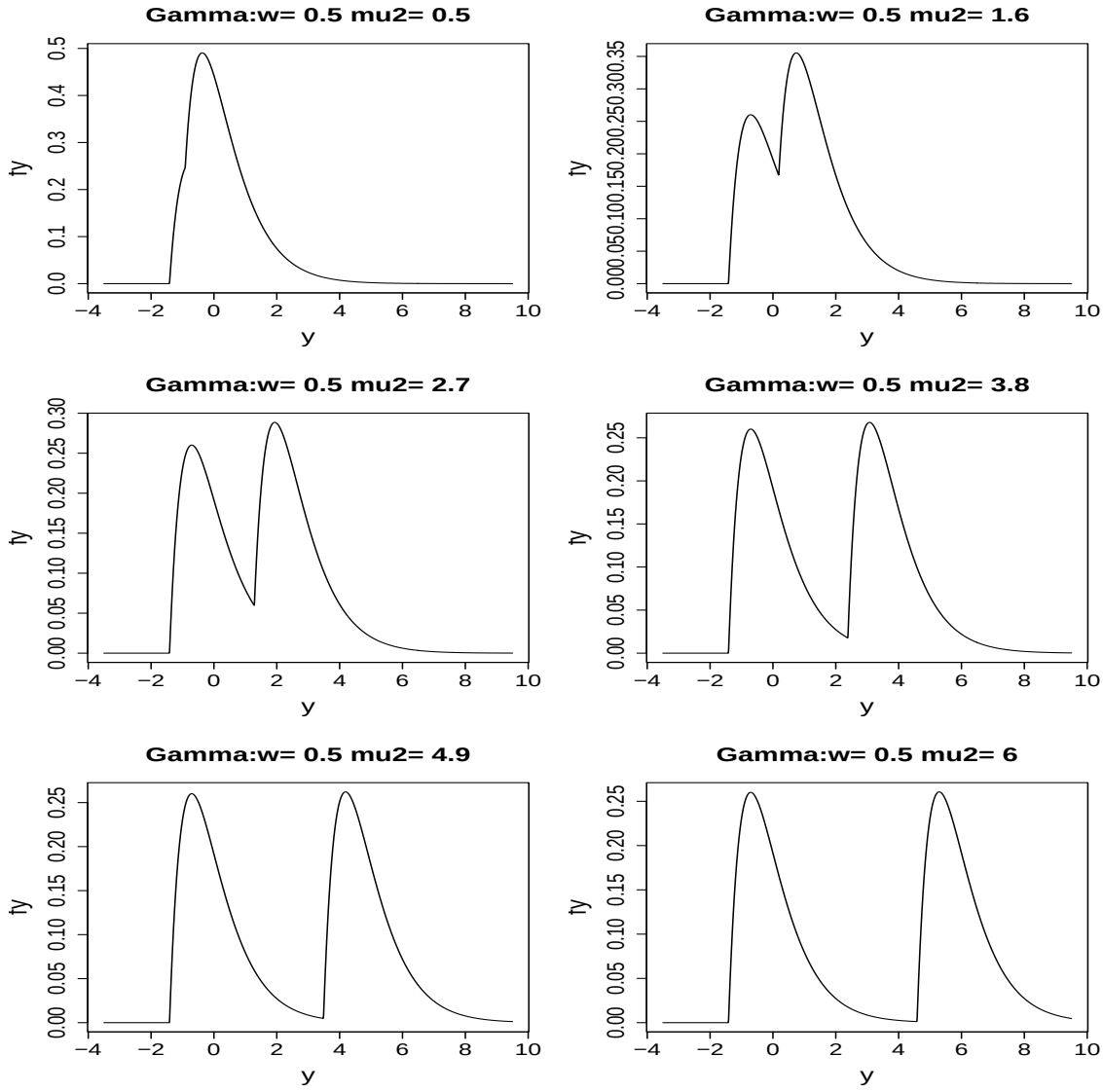

Figure 12: Mixture of Gamma Densities : $w_1 = 0.5$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

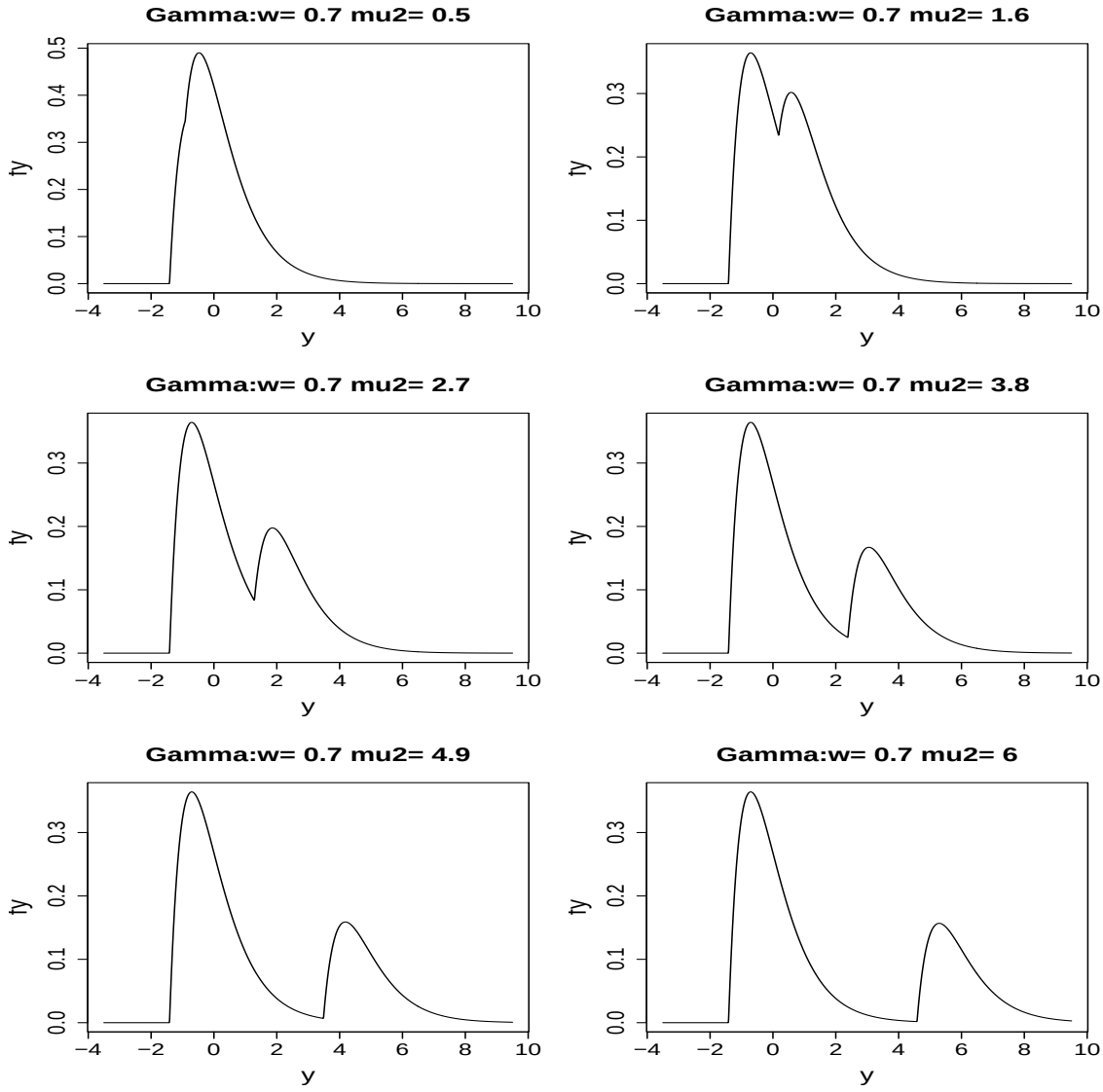

Figure 13: Mixture of Gamma Densities : $w_1 = 0.7$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

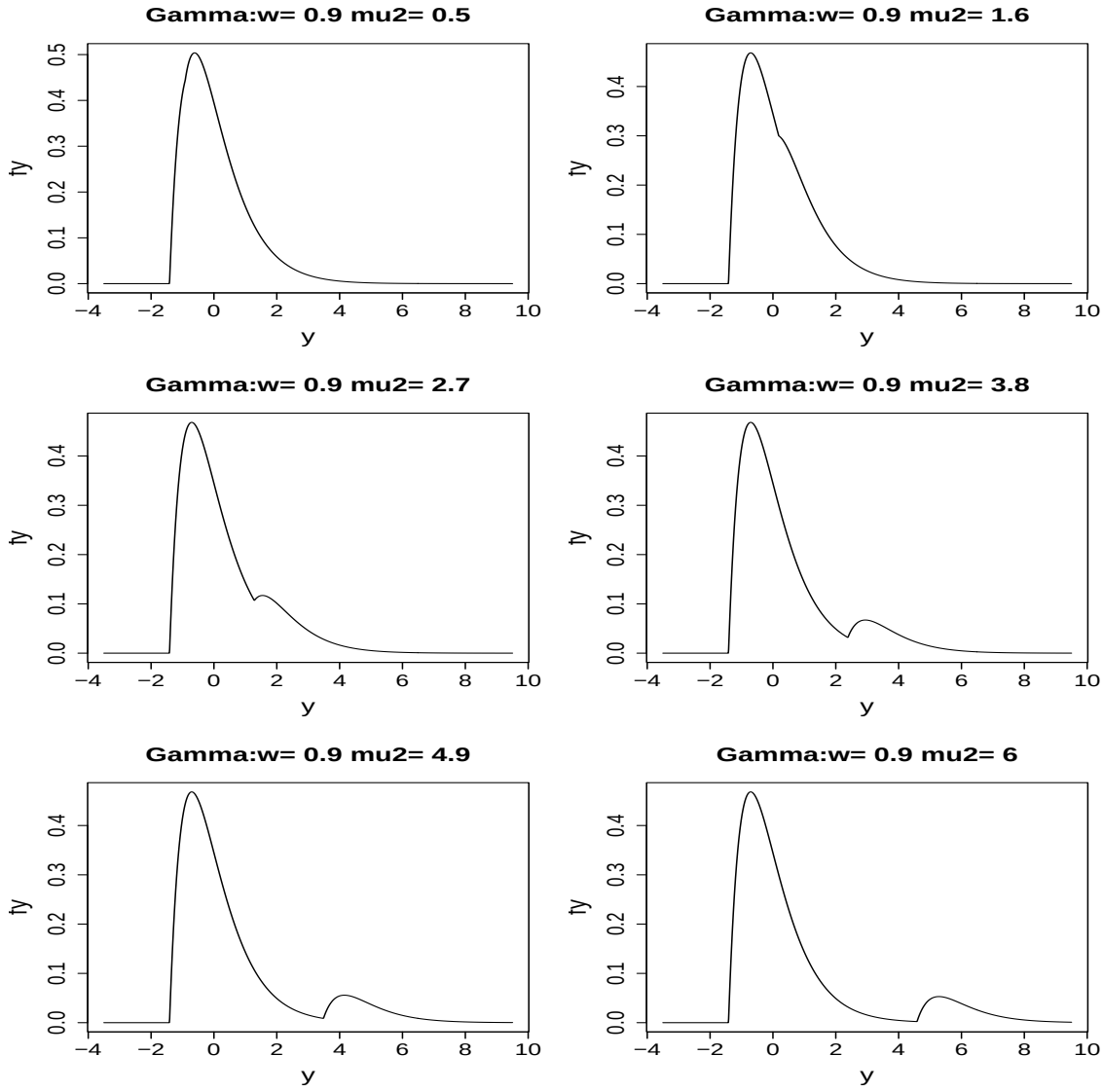

Figure 14: Mixture of Gamma Densities : $w_1 = 0.9$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

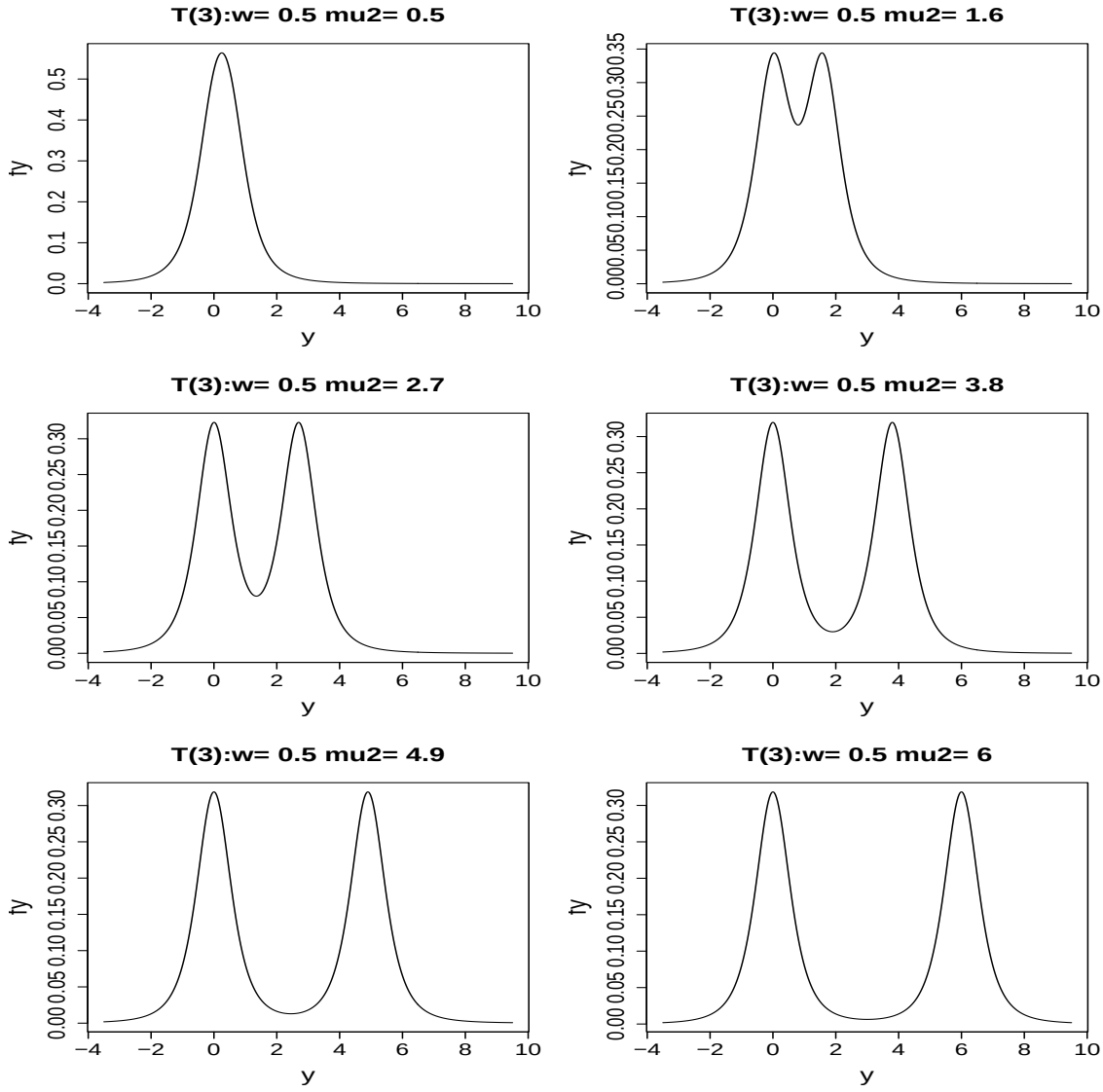

Figure 15: Mixture of T_3 Densities : $w_1 = 0.5$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

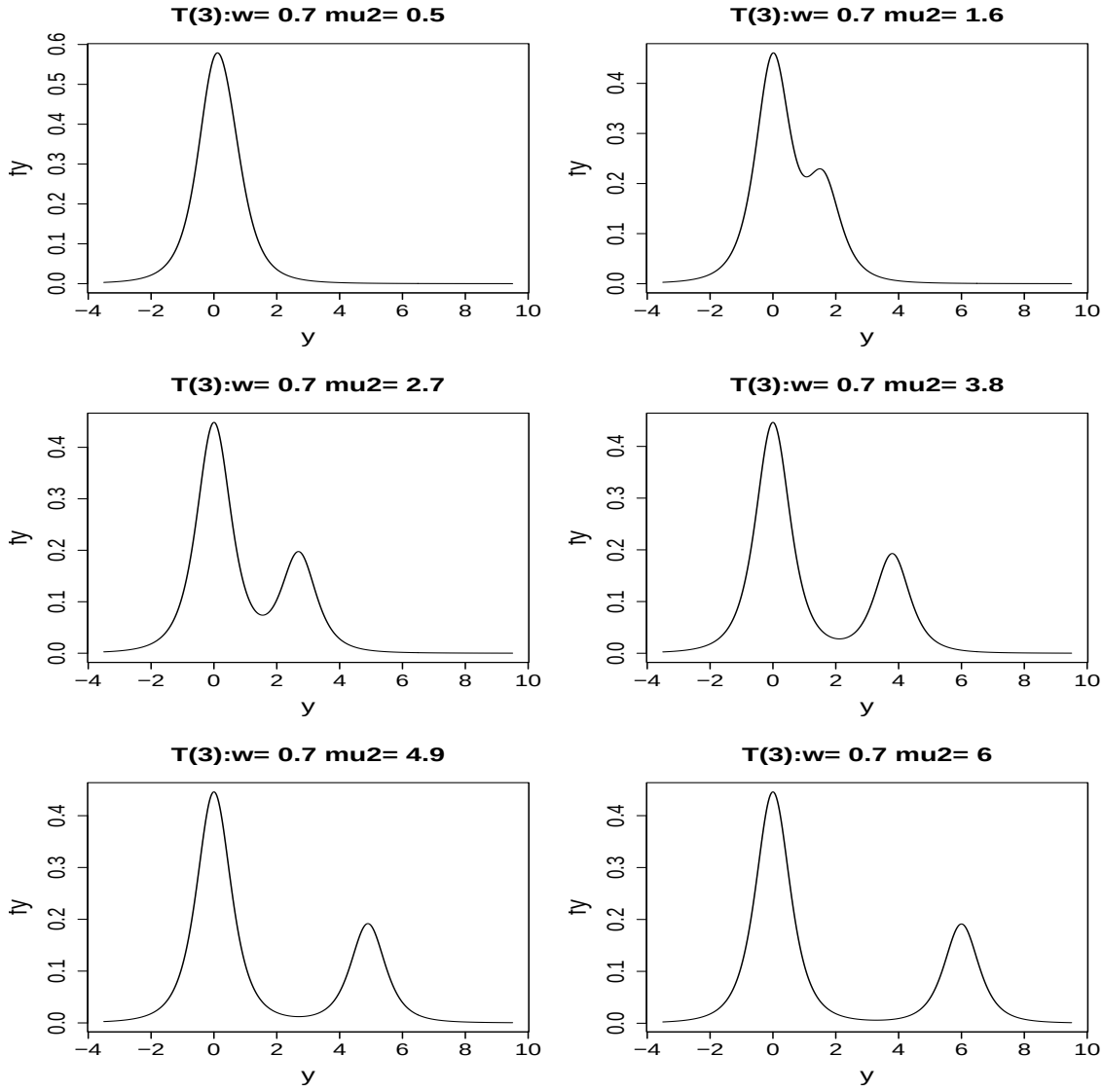

Figure 16: Mixture of T_3 Densities : $w_1 = 0.7$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

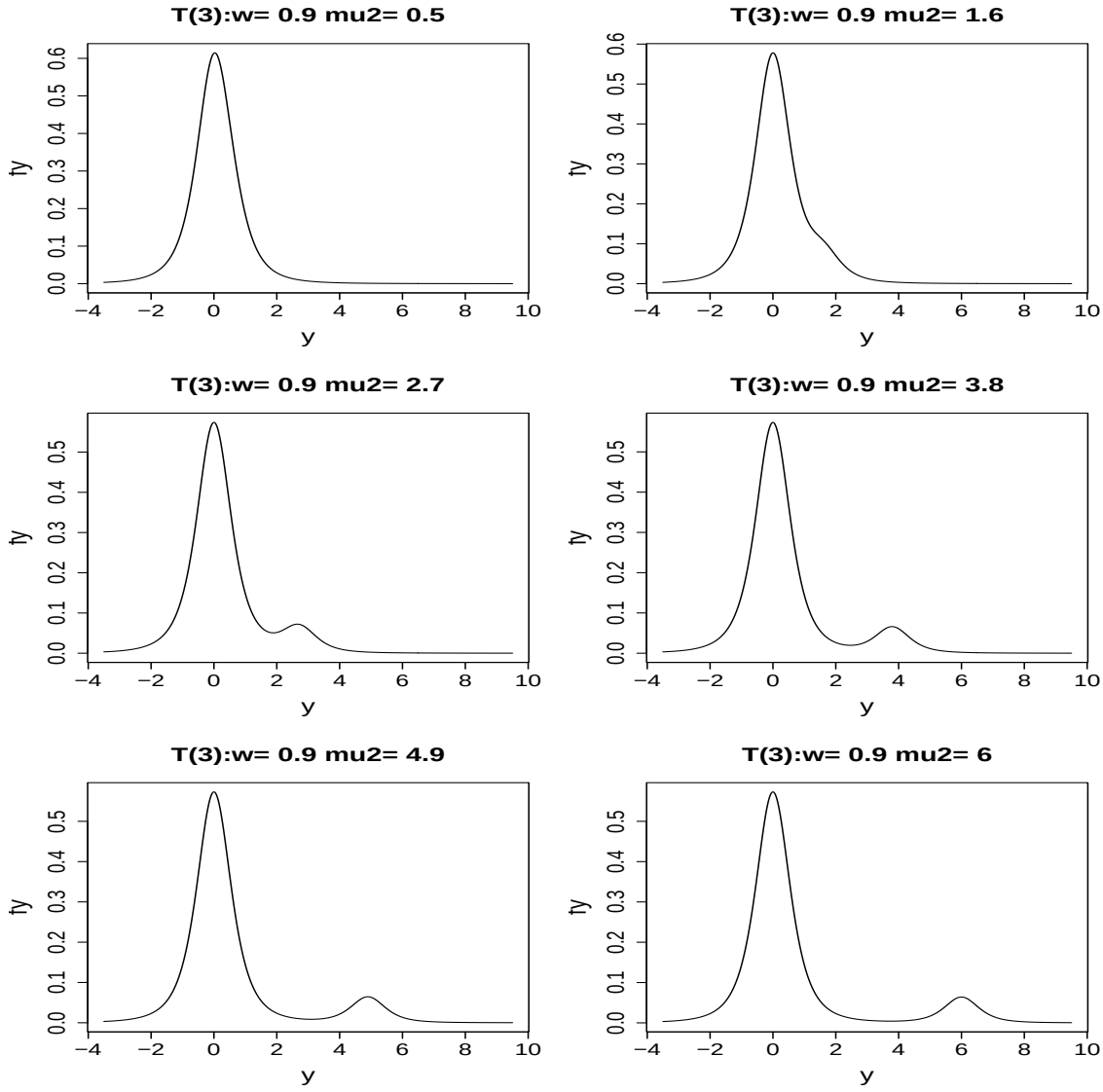

Figure 17: Mixture of T_3 Densities : $w_1 = 0.9$, $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$

APPENDIX B

ITERATIONS AS DEFINED BY (2.13) FOR LAPLACE MIXTURE

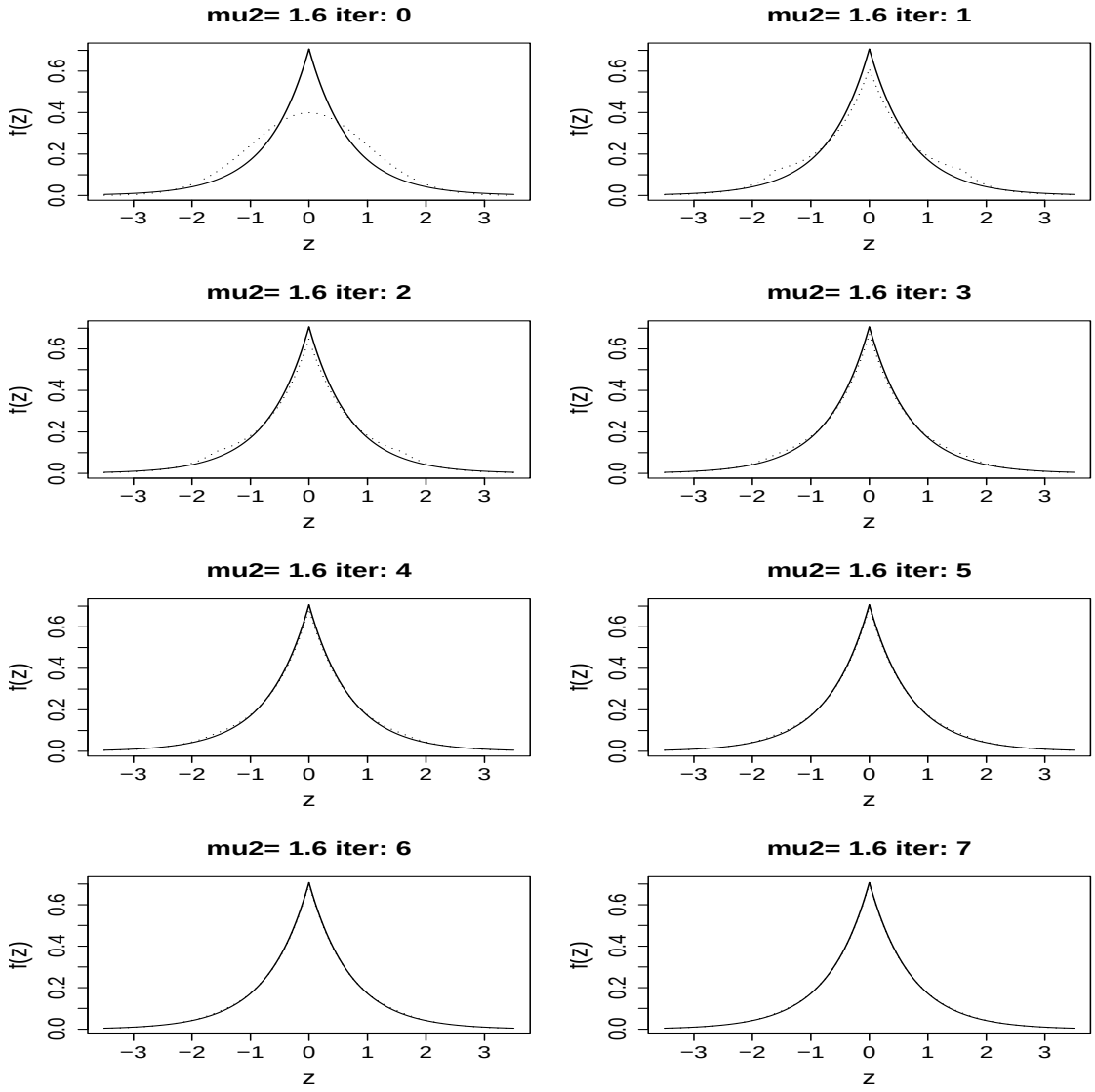

Figure 18: f_r for Laplace : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 1.6$, $\sigma_1 = \sigma_2 = 1$

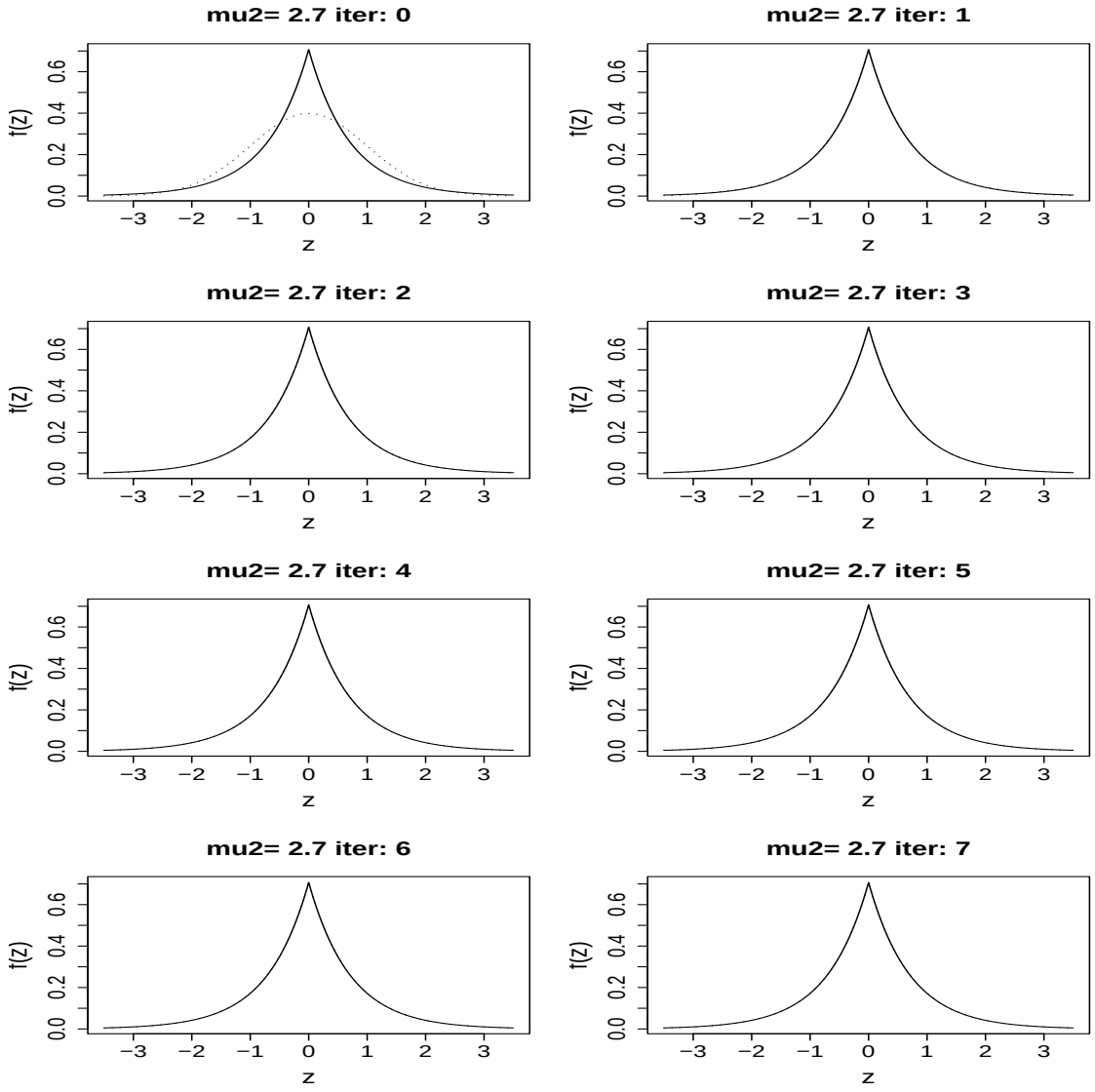

Figure 19: f_r for Laplace : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 2.7$, $\sigma_1 = \sigma_2 = 1$

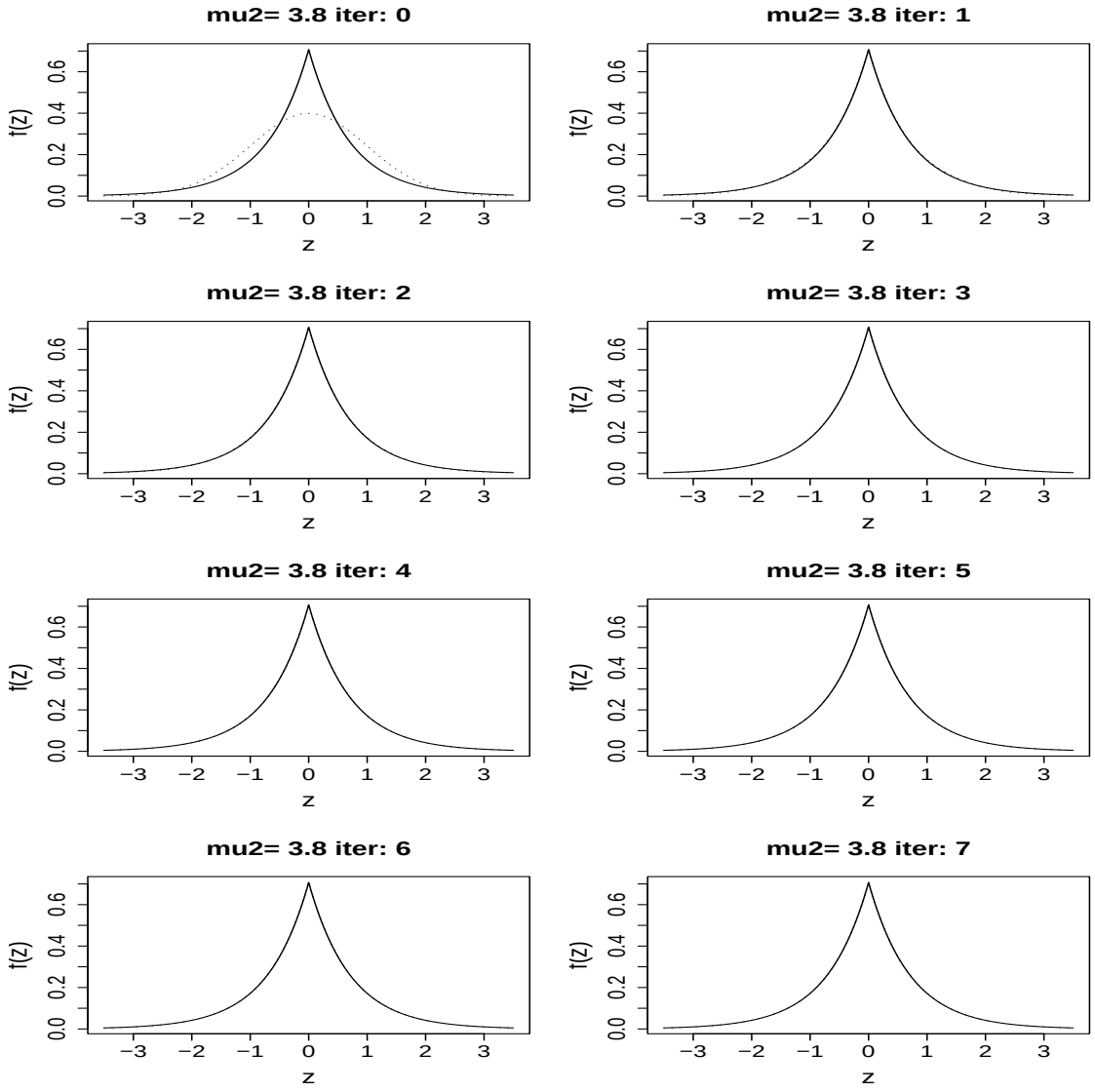

Figure 20: f_r for Laplace : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 3.8$, $\sigma_1 = \sigma_2 = 1$

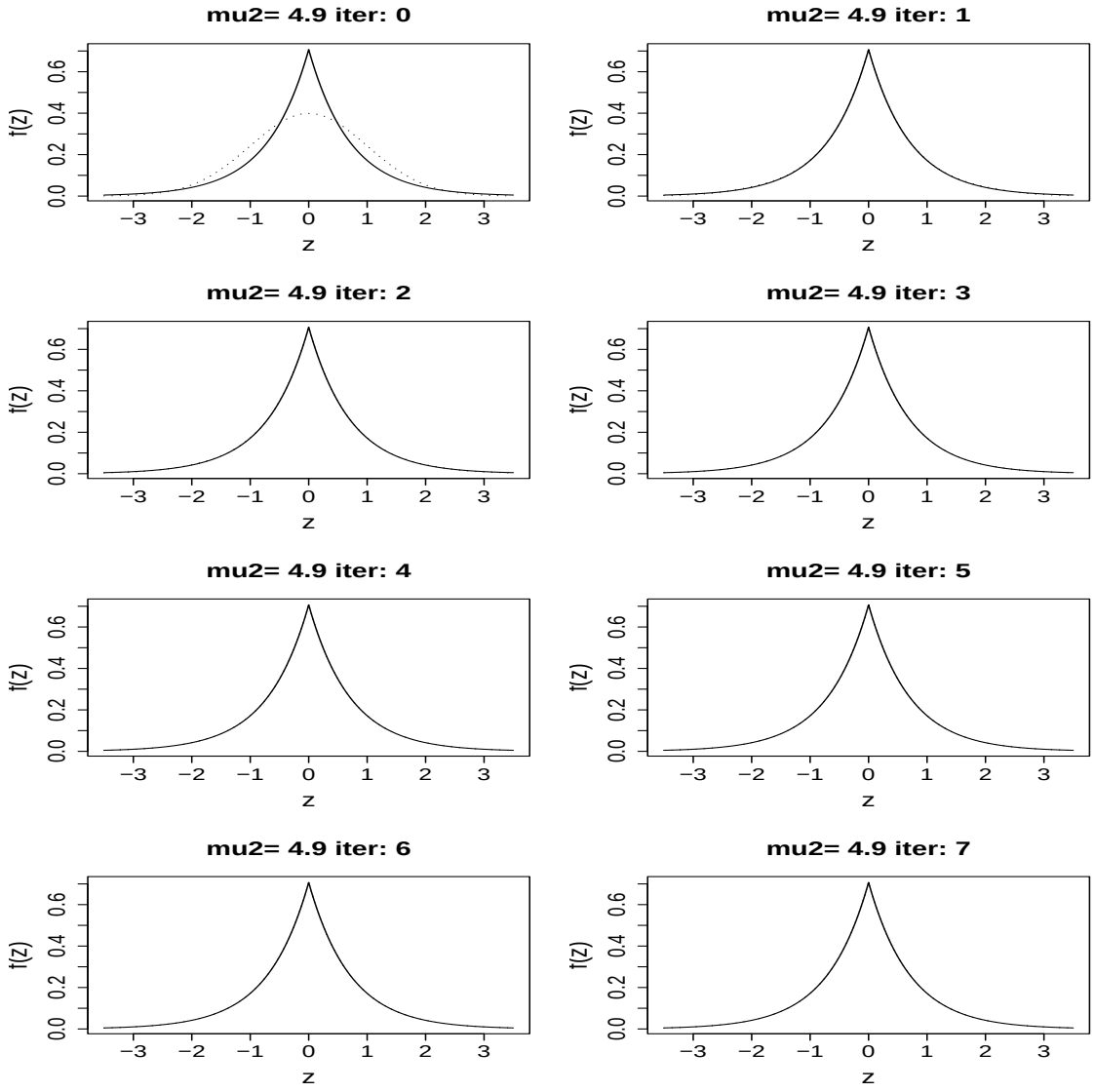

Figure 21: f_r for Laplace : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 4.9$, $\sigma_1 = \sigma_2 = 1$

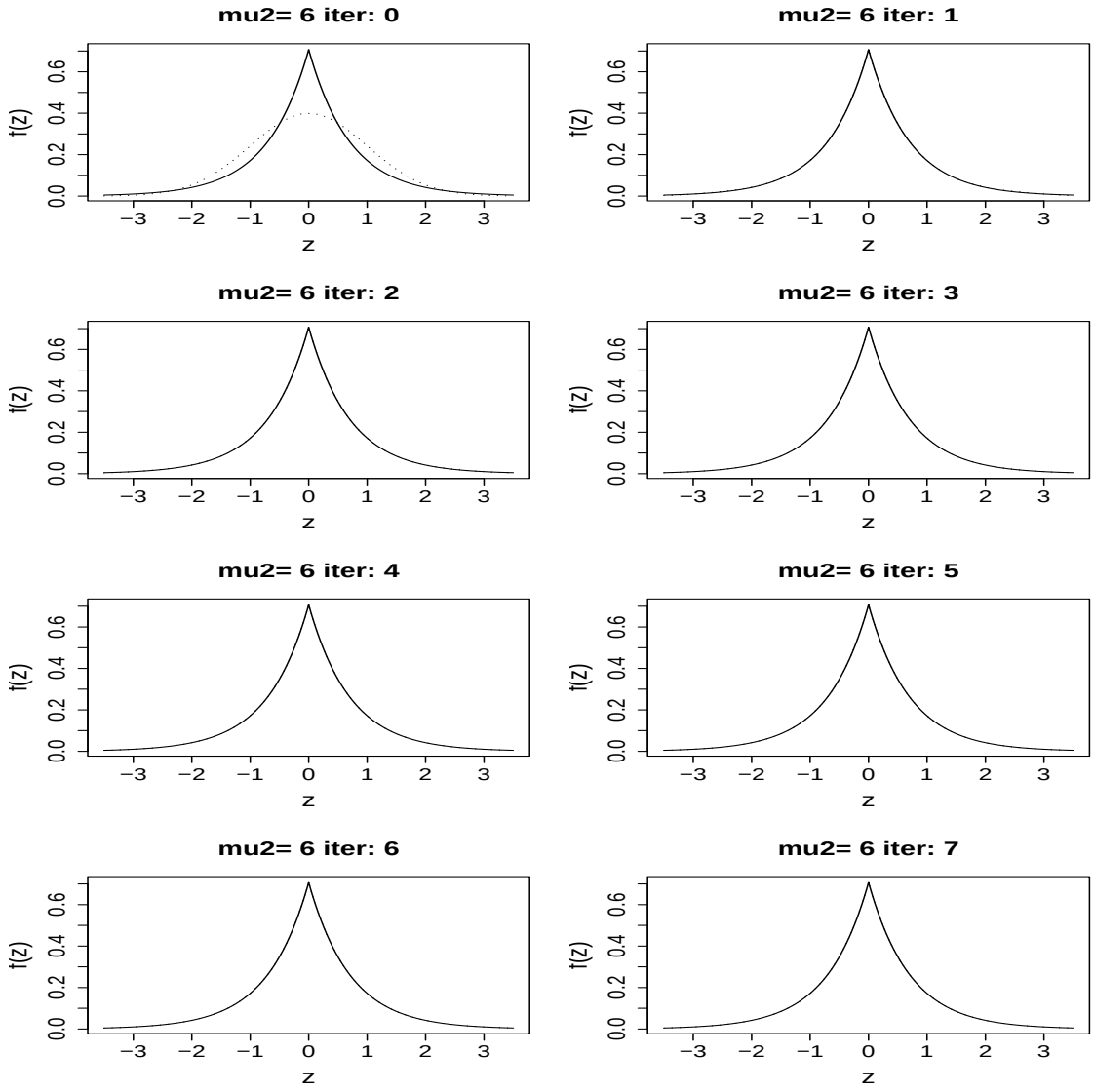

Figure 22: f_r for Laplace : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 6.0$, $\sigma_1 = \sigma_2 = 1$

APPENDIX C

ITERATIONS AS DEFINED BY (2.13) FOR GAMMA MIXTURE

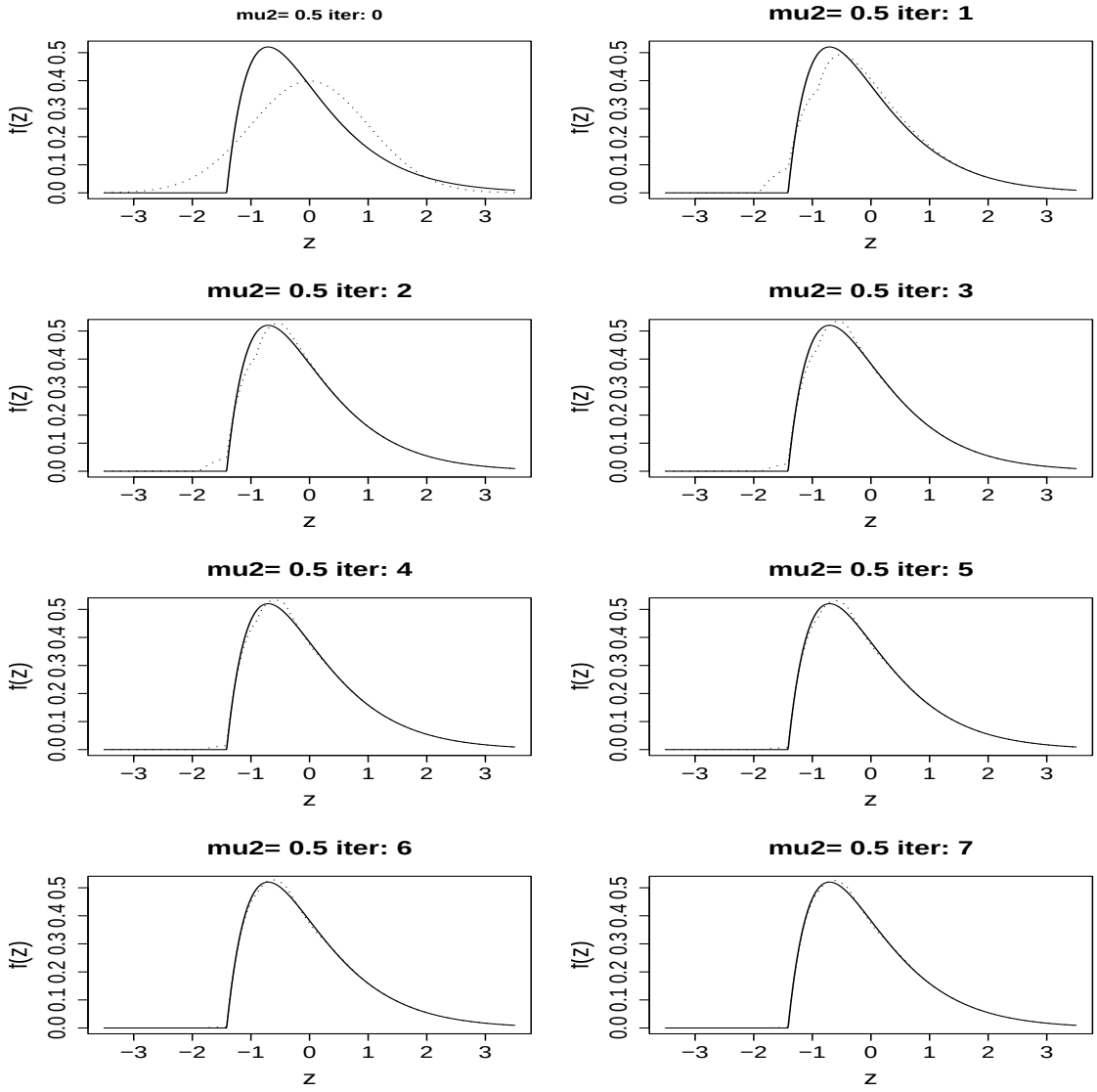

Figure 23: f_r for Gamma : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$

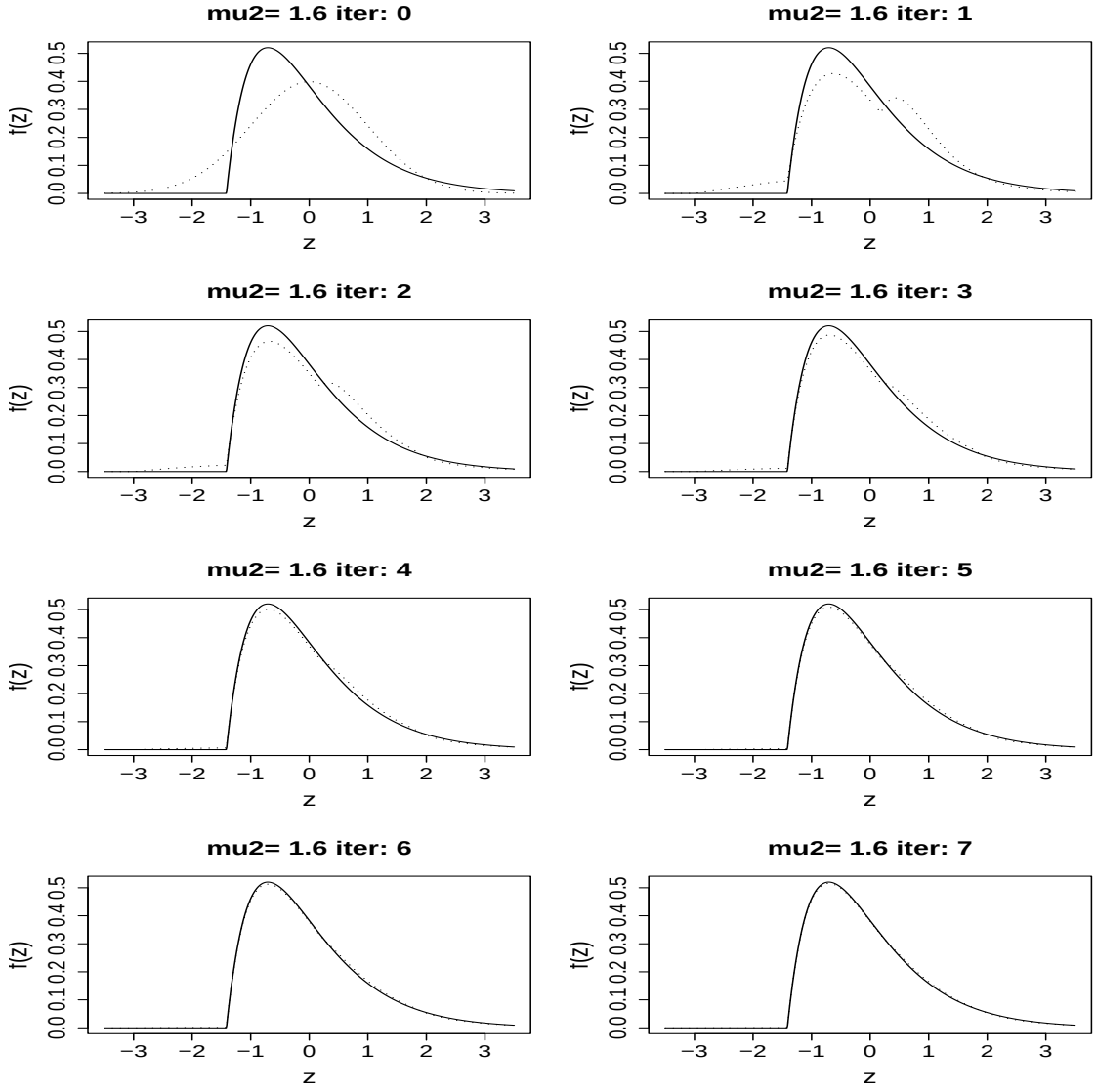

Figure 24: f_r for Gamma : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 1.6$, $\sigma_1 = \sigma_2 = 1$

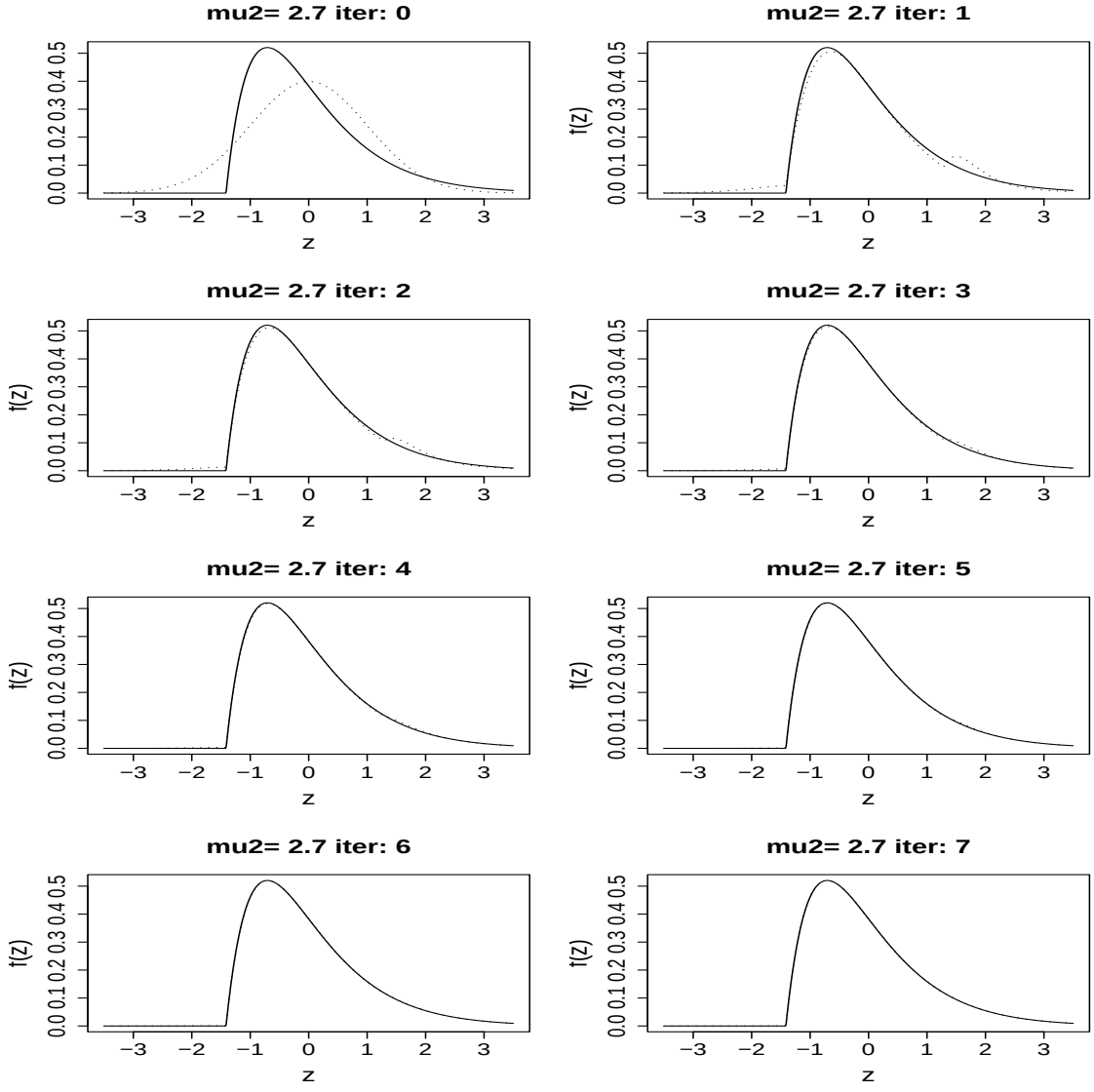

Figure 25: f_r for Gamma : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 2.7$, $\sigma_1 = \sigma_2 = 1$

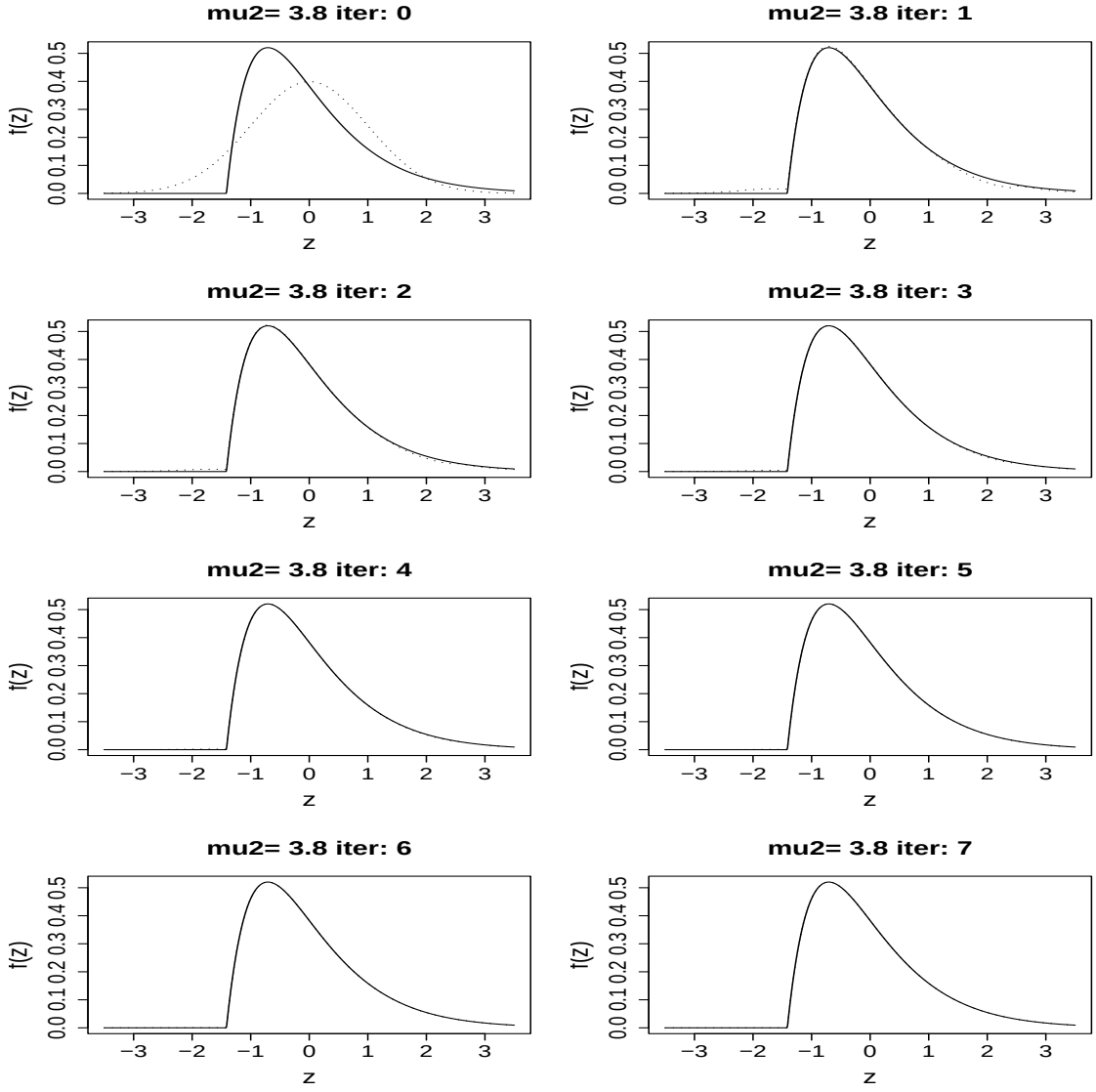

Figure 26: f_r for Gamma : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 3.8$, $\sigma_1 = \sigma_2 = 1$

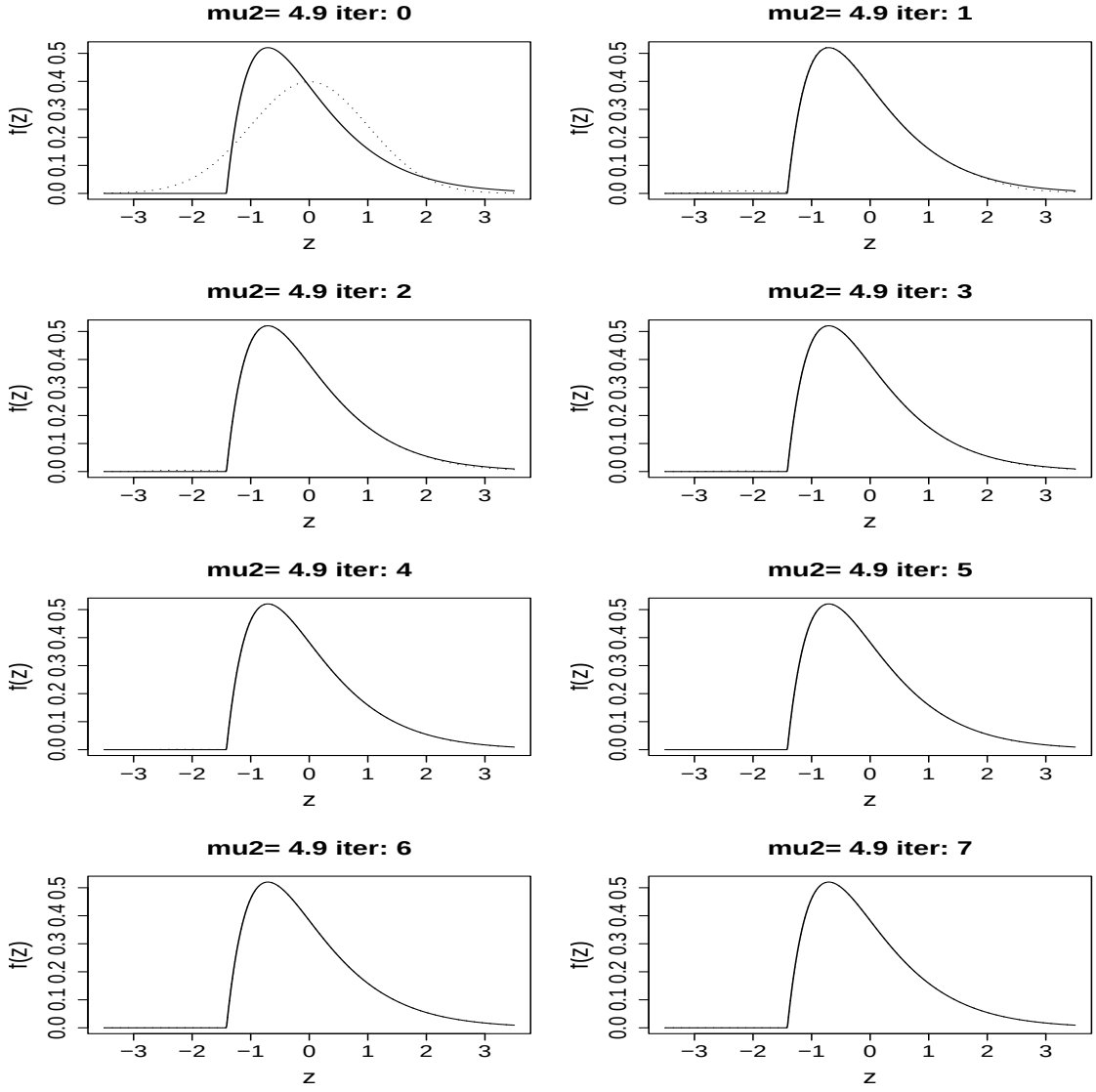

Figure 27: f_r for Gamma : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 4.9$, $\sigma_1 = \sigma_2 = 1$

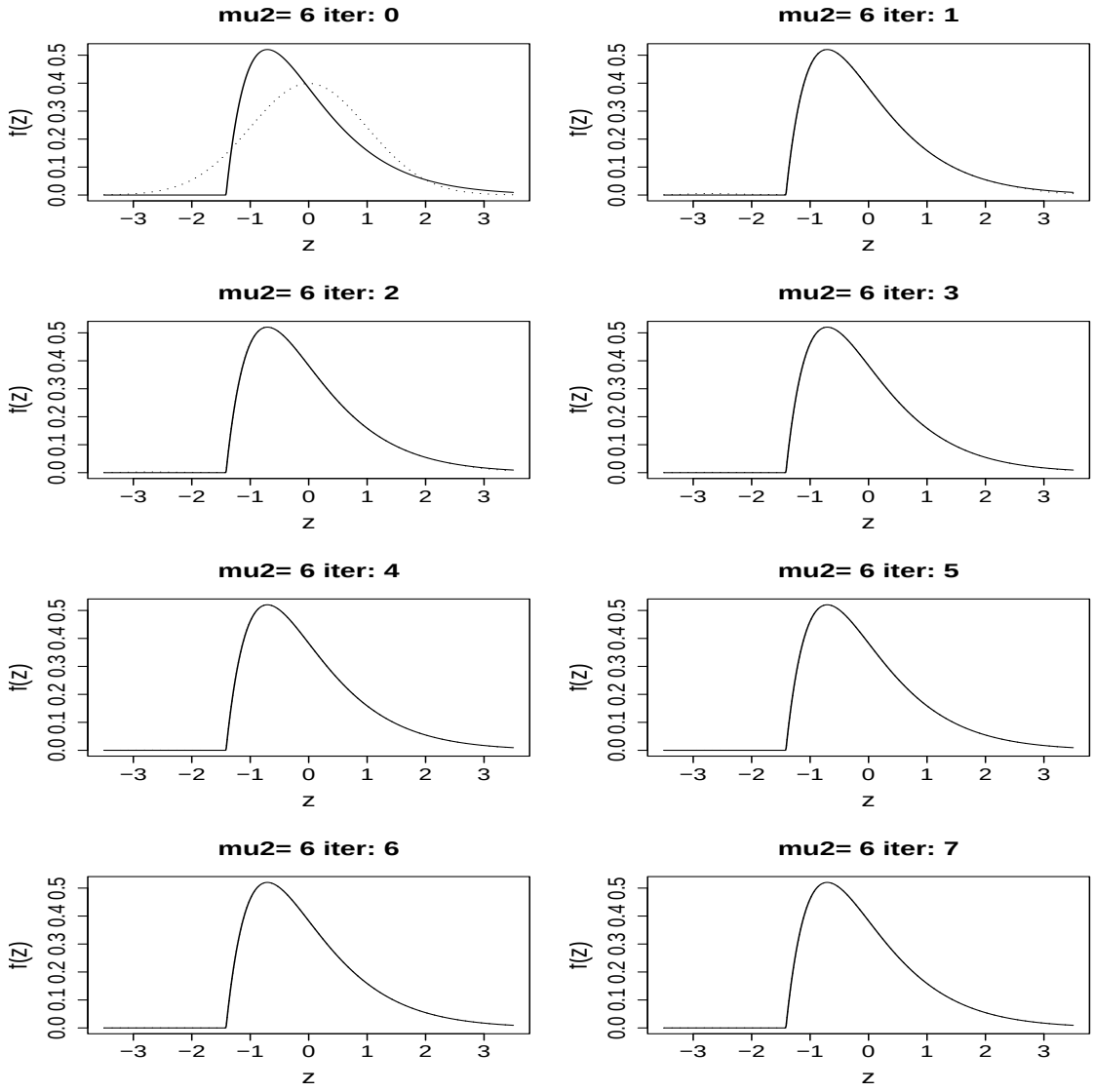

Figure 28: f_r for Gamma : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 6.0$, $\sigma_1 = \sigma_2 = 1$

APPENDIX D

ITERATIONS AS DEFINED BY (2.13) FOR T_3 MIXTURE

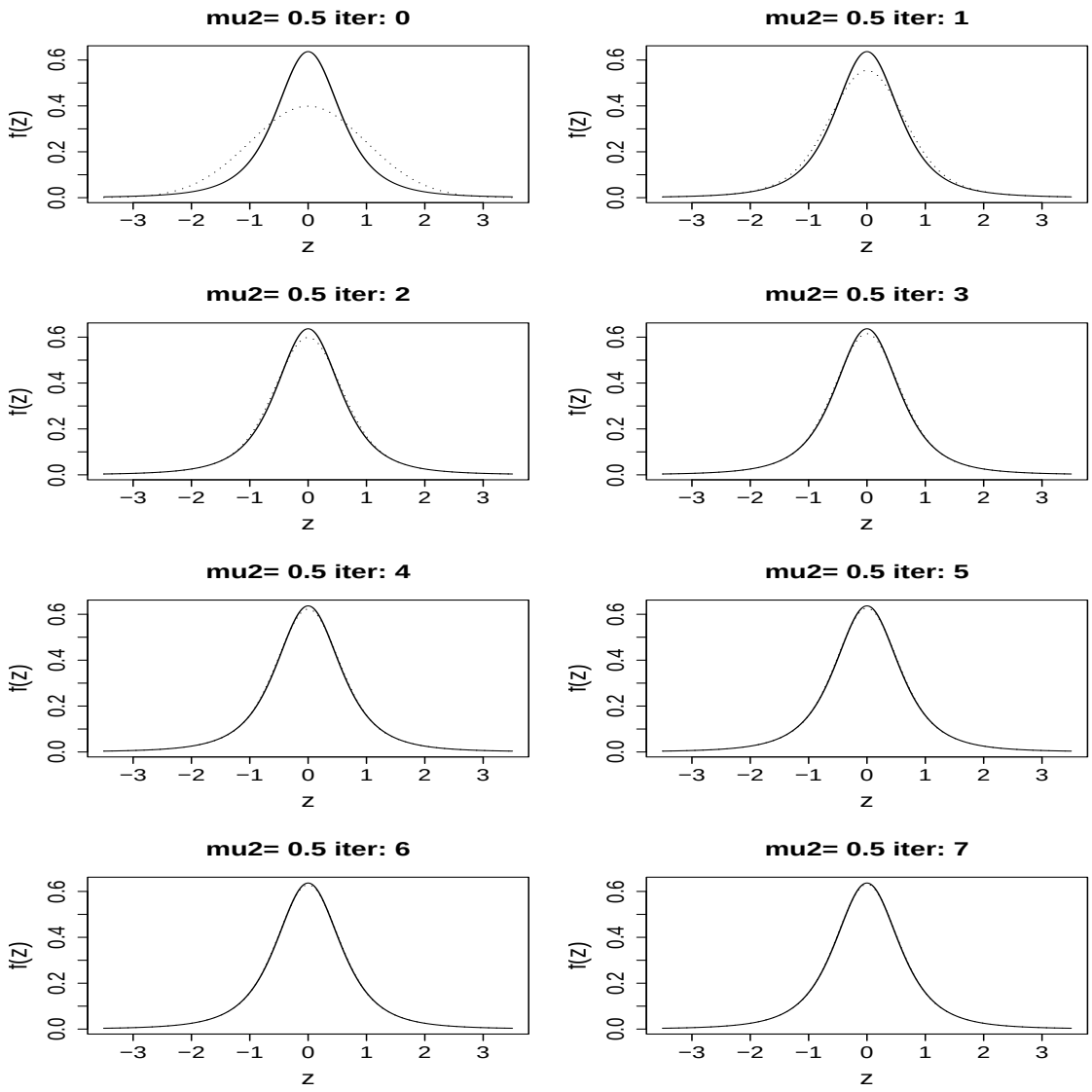

Figure 29: f_r for T_3 : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$

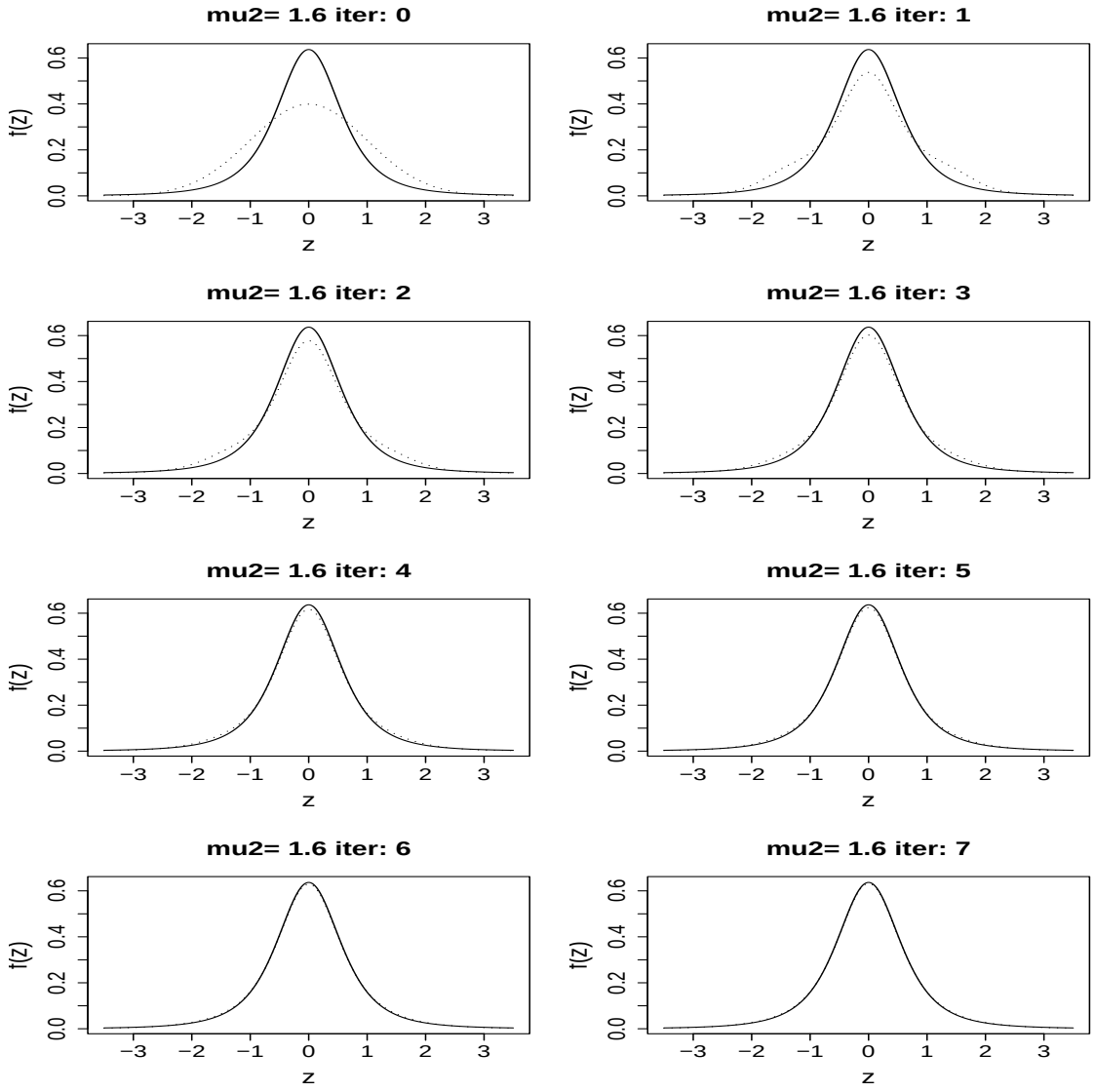

Figure 30: f_r for T_3 : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 1.6$, $\sigma_1 = \sigma_2 = 1$

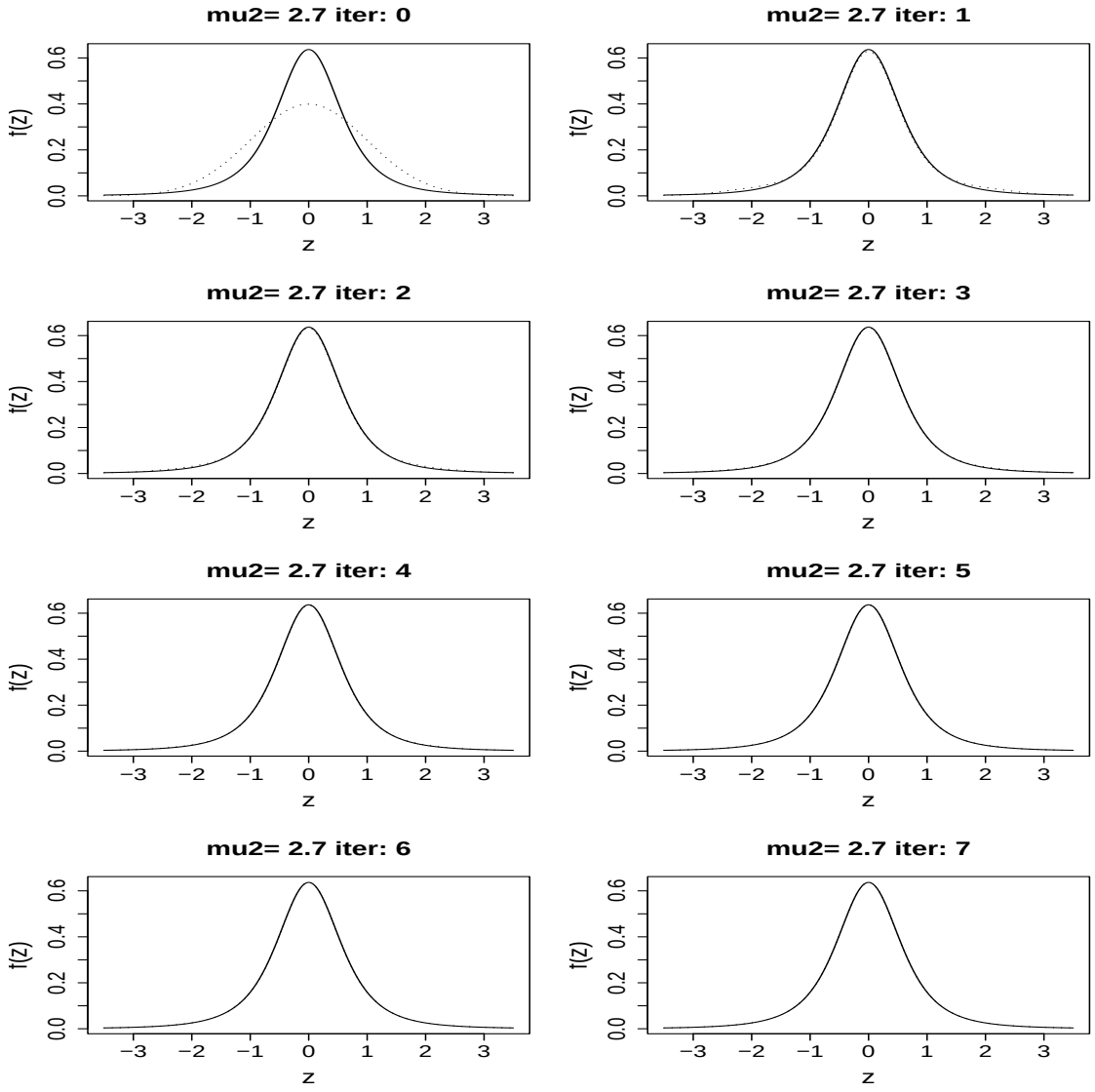

Figure 31: f_r for T_3 : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 2.7$, $\sigma_1 = \sigma_2 = 1$

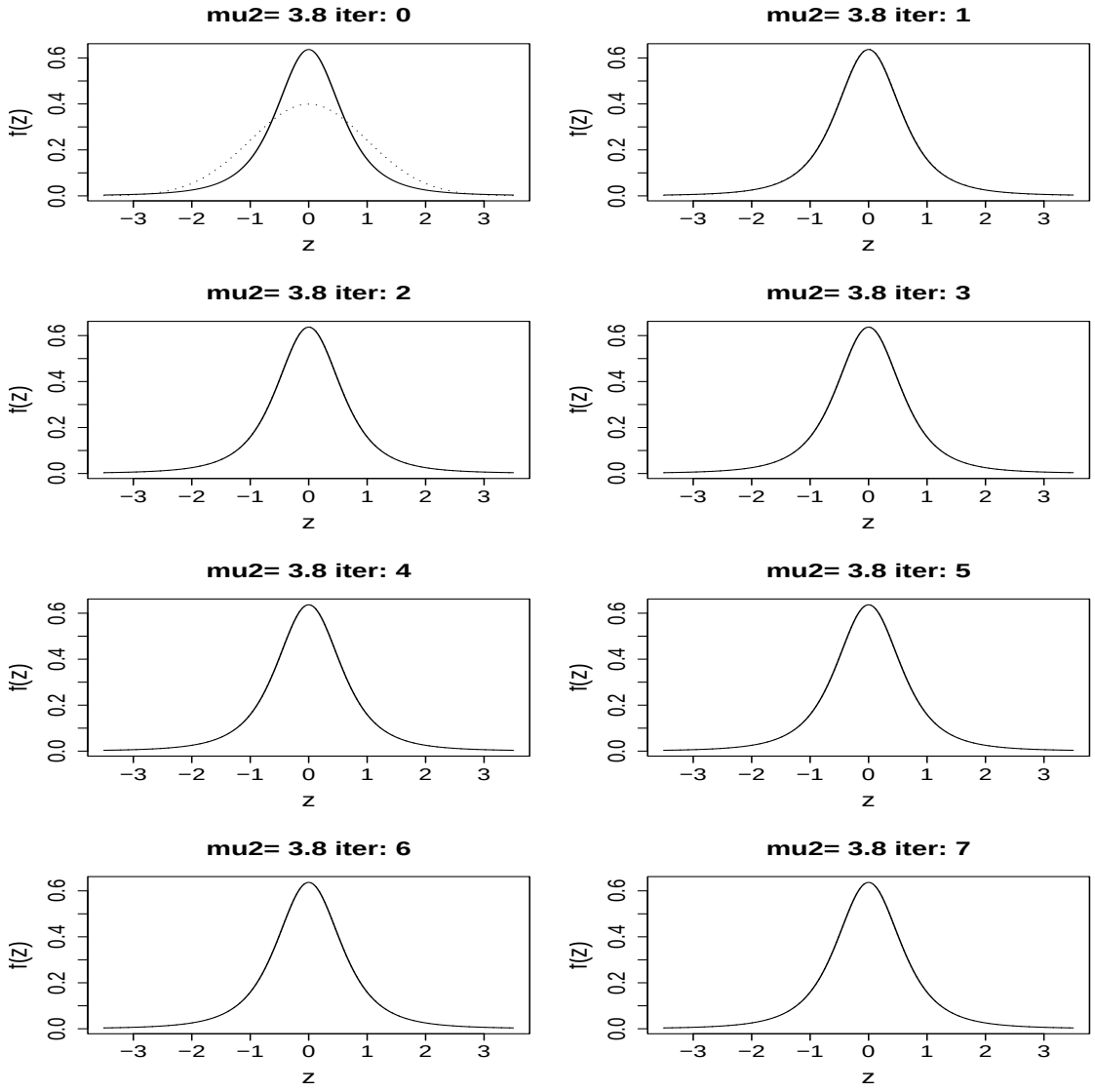

Figure 32: f_r for T_3 : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 3.8$, $\sigma_1 = \sigma_2 = 1$

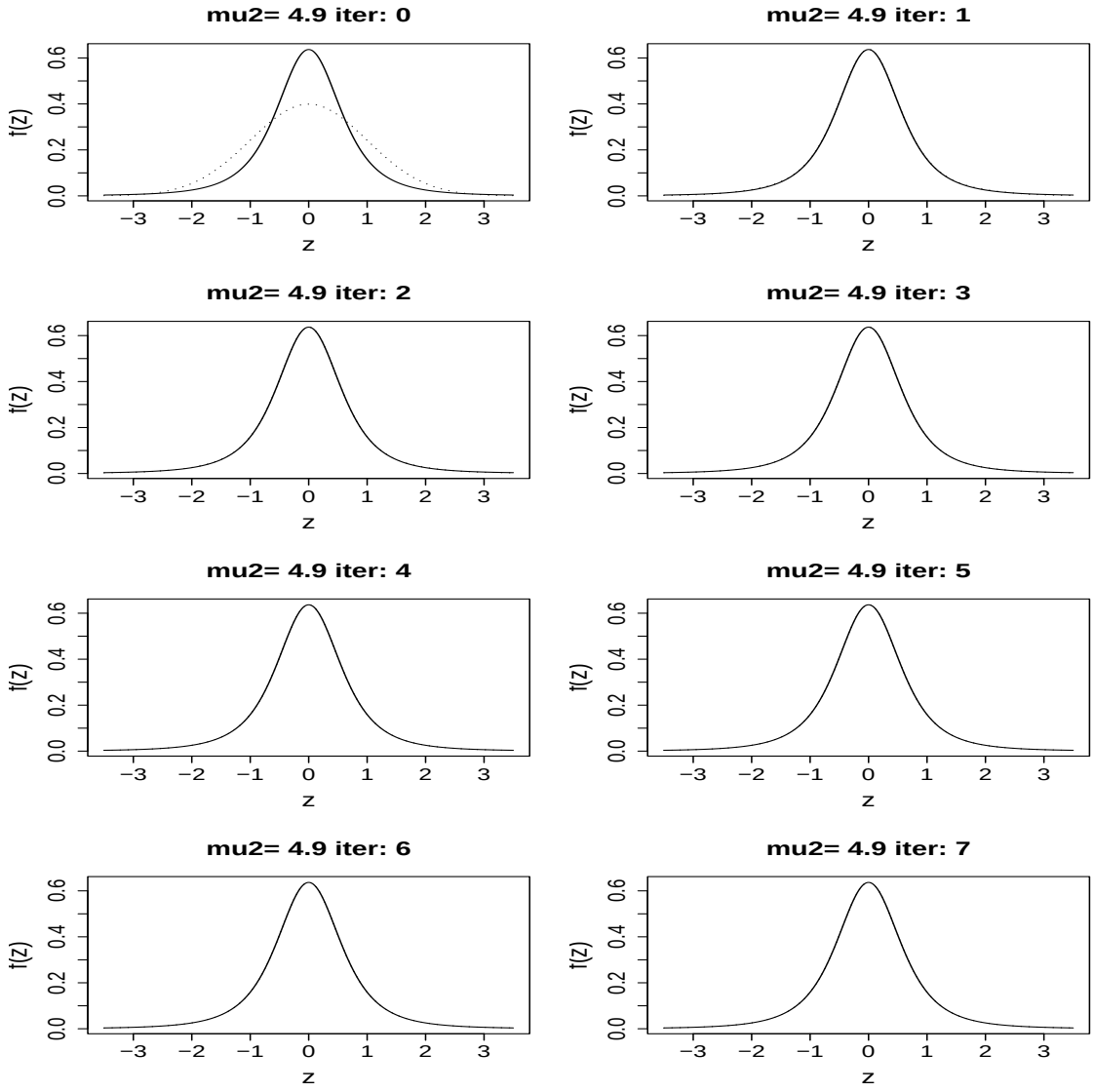

Figure 33: f_r for T_3 : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 4.9$, $\sigma_1 = \sigma_2 = 1$

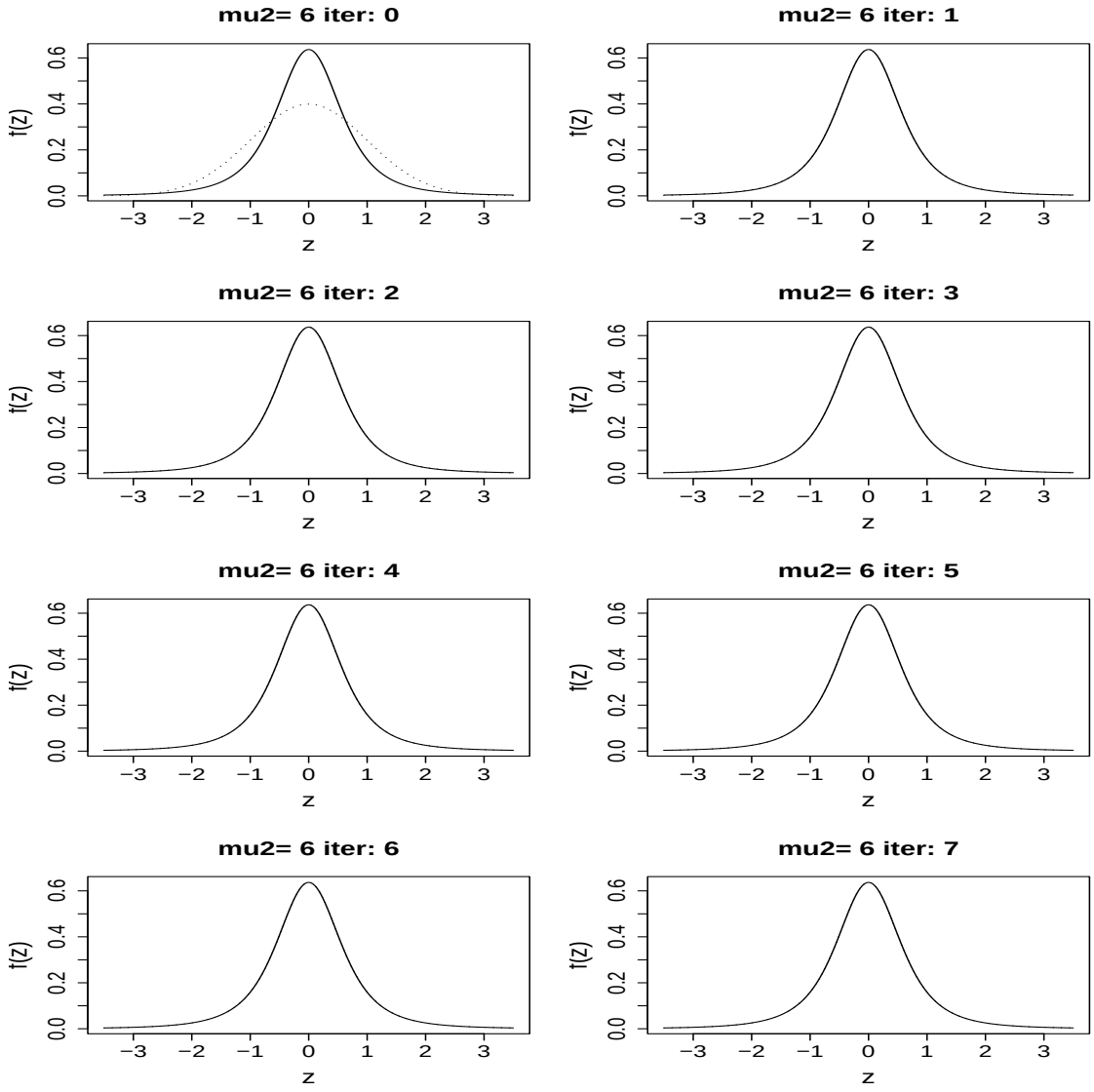

Figure 34: f_r for T_3 : $w_1 = 0.5$, $\mu_1 = 0$, $\mu_2 = 6.0$, $\sigma_1 = \sigma_2 = 1$

APPENDIX E

EMPIRICAL DISTRIBUTIONS OF P-VALUES

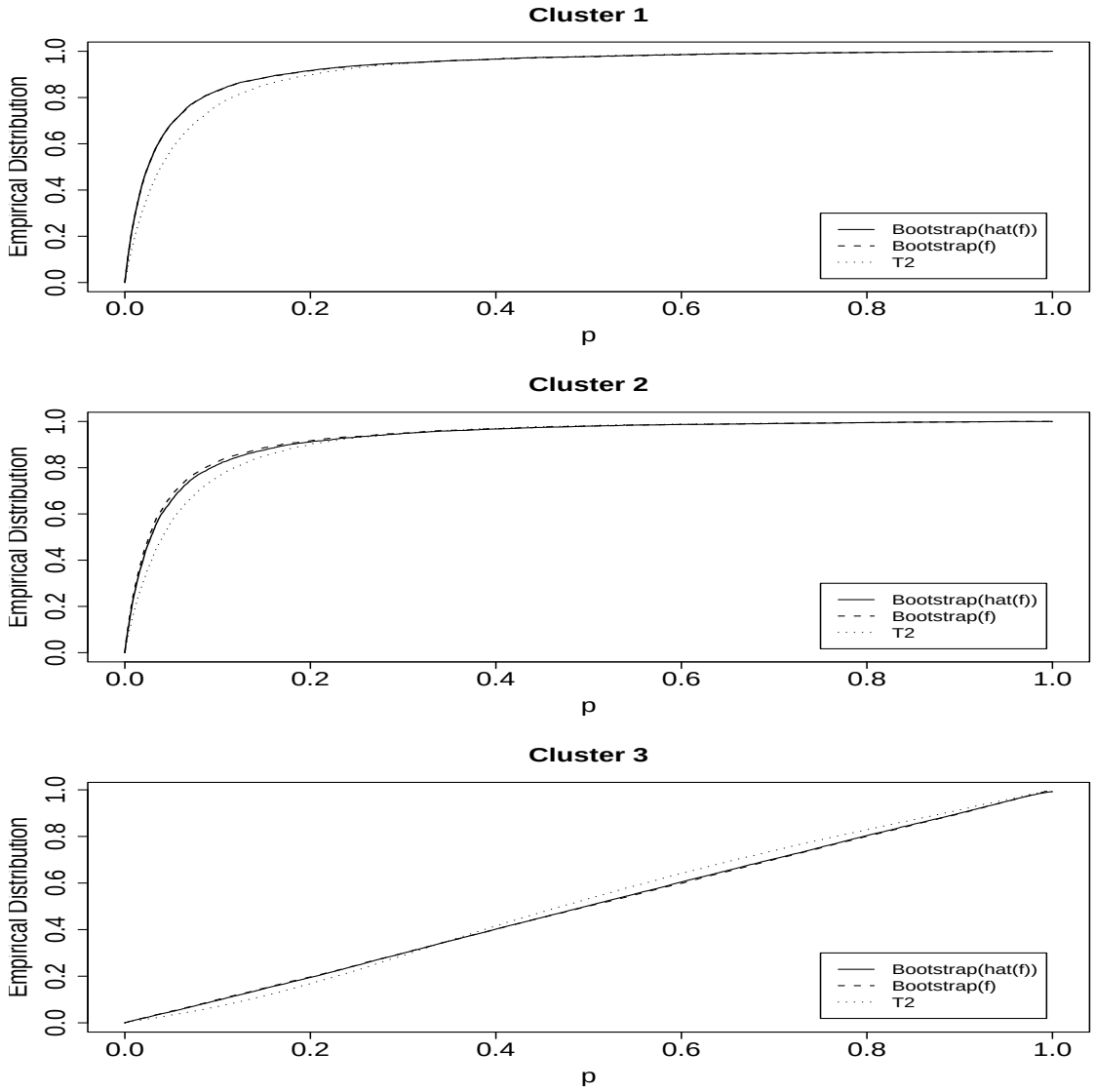

Figure 35: Empirical Distribution of p-values for Case 2 of Laplace Mixture

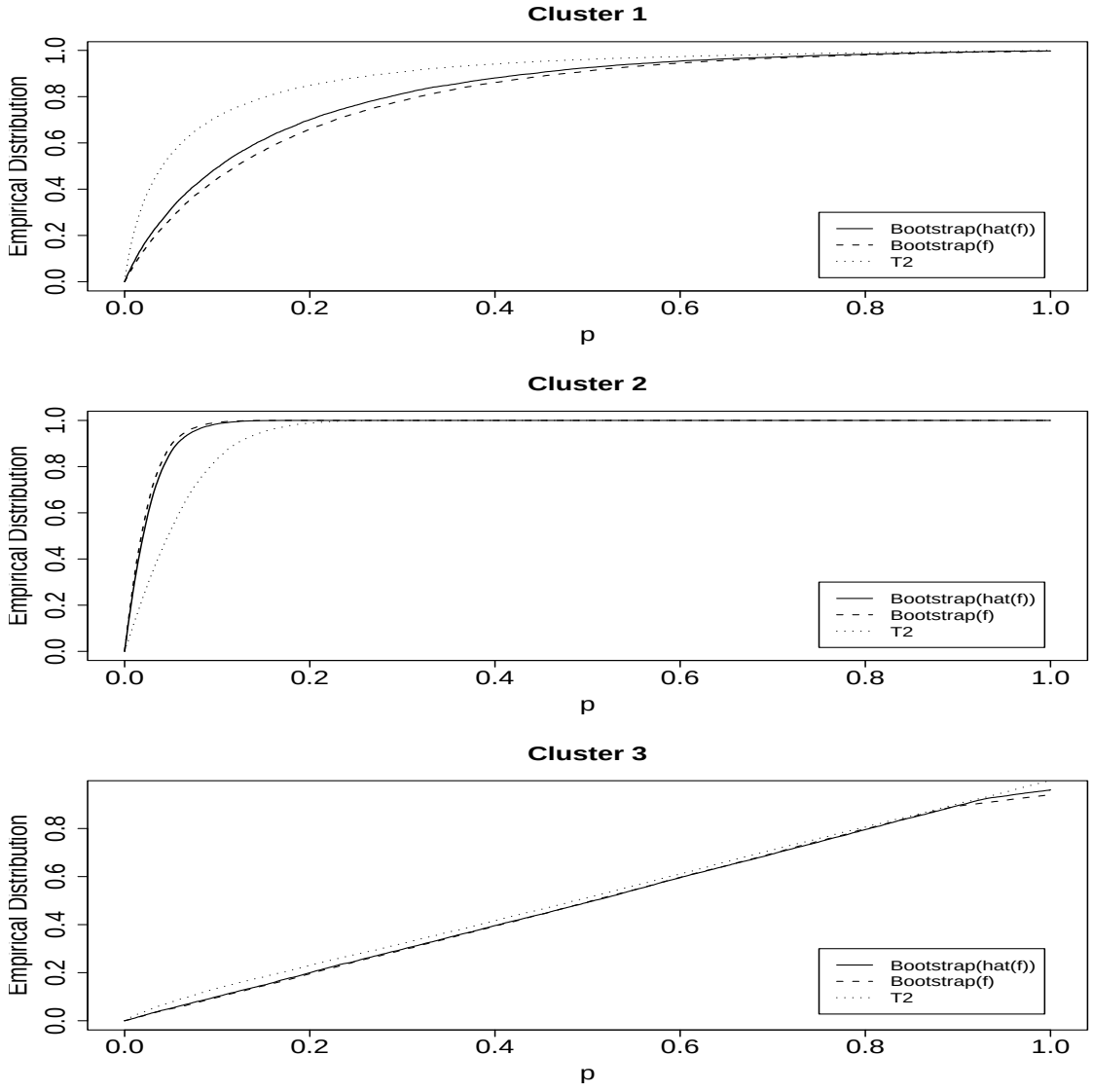

Figure 36: Empirical Distribution of p-values for Case 1 of Gamma Mixture

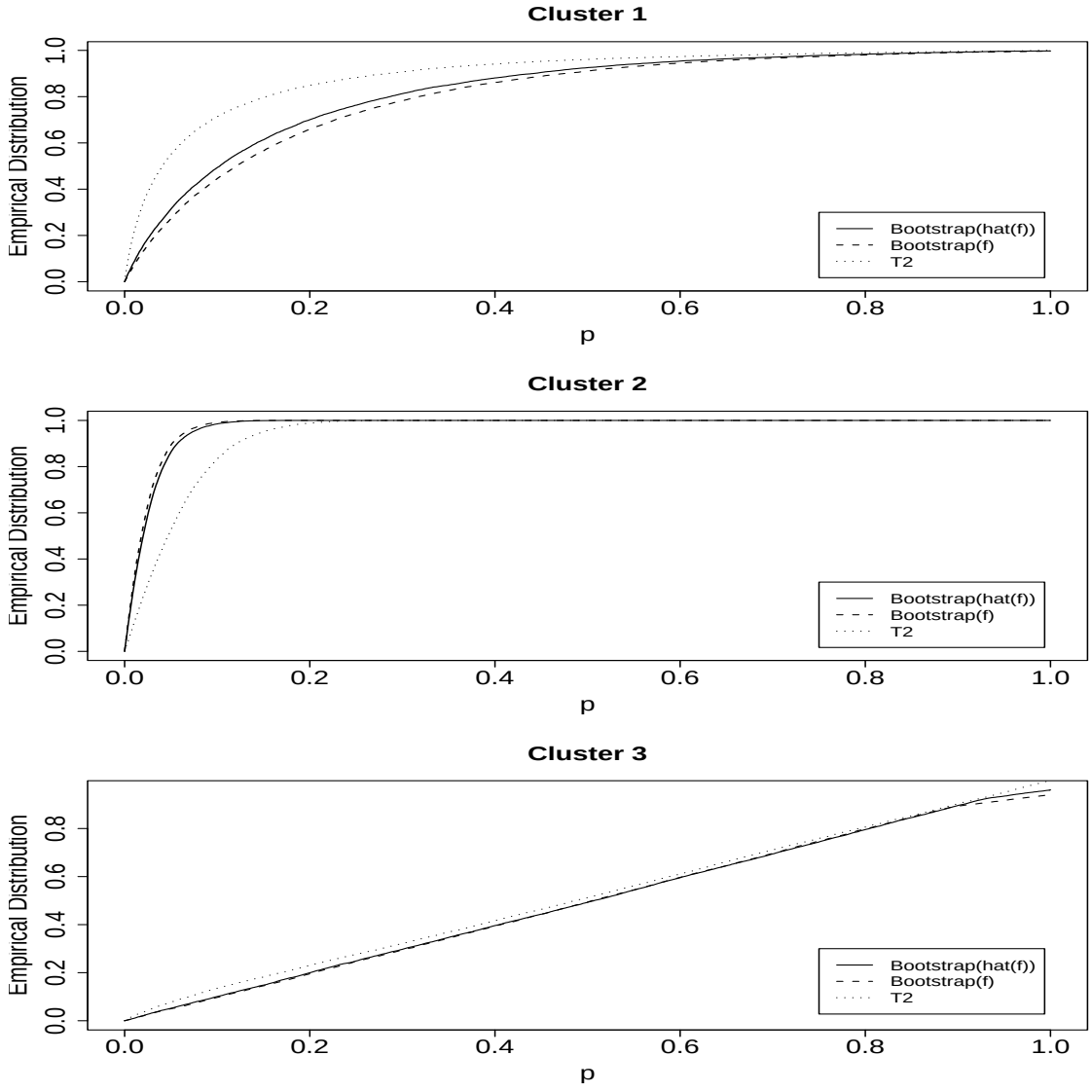

Figure 37: Empirical Distribution of p-values for Case 2 of Gamma Mixture

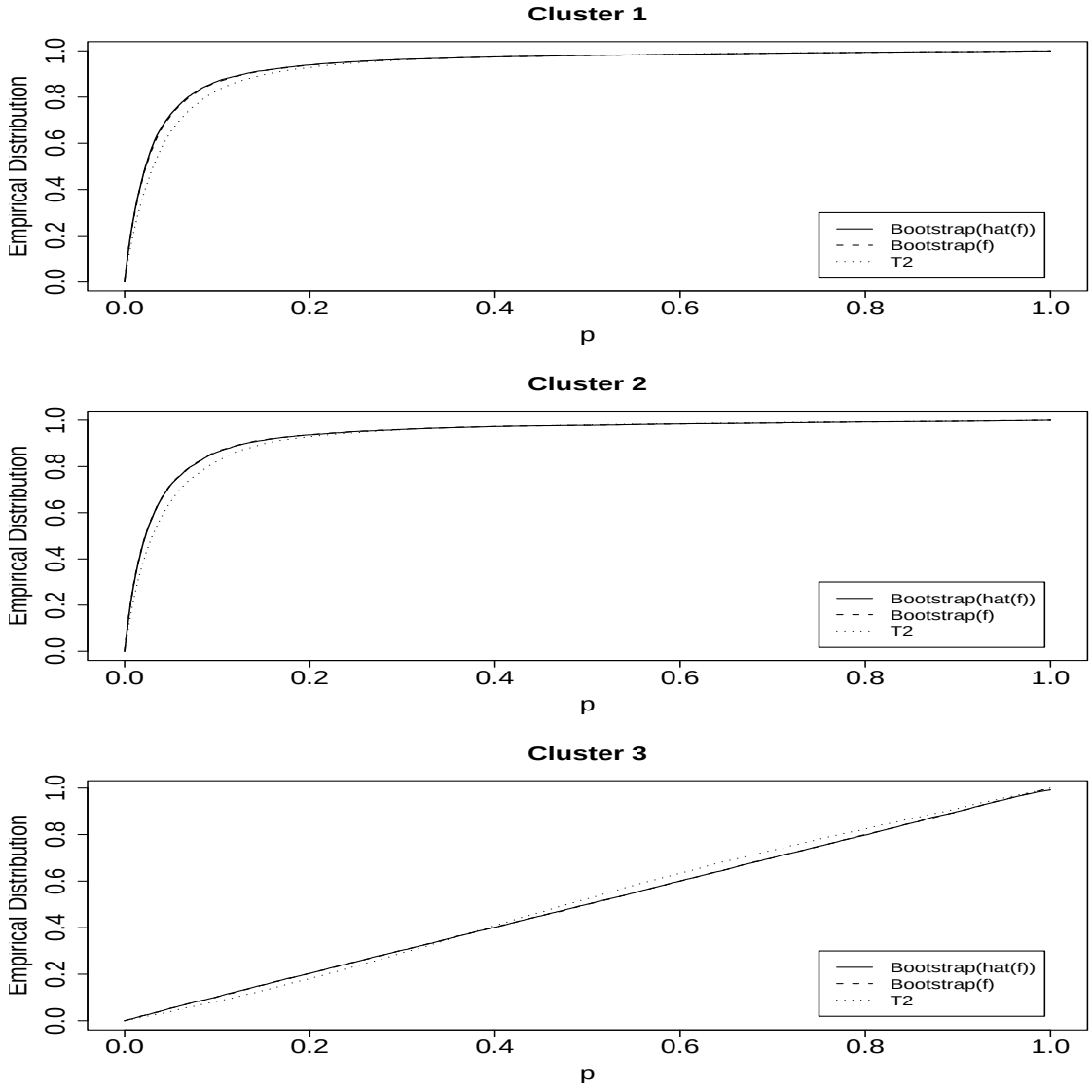

Figure 38: Empirical Distribution of p-values for Case 1 of T3 Mixture

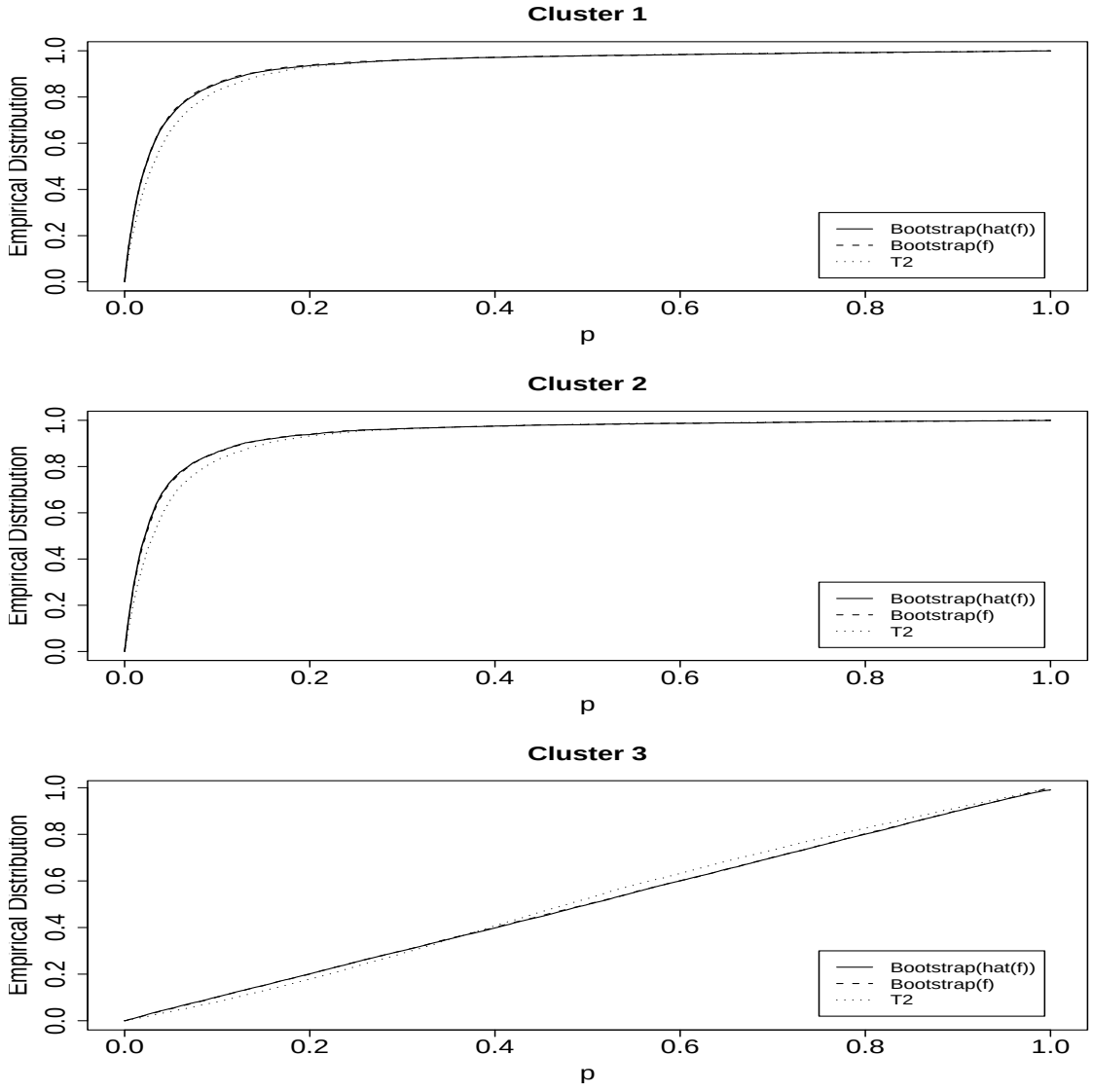

Figure 39: Empirical Distribution of p-values for Case 2 of T3 Mixture

VITA

Juhee Song was born in Booksahm, Korea on November 14, 1972. She graduated from Deokwon Girls High School in Seoul, Korea in 1991. She received a Bachelor of Science in statistics from Inha University in 1995. Later, she received a Master of Art in applied statistics from Yonsei University in 1998. She started her study in statistics at Texas A&M University in 1999 and received her Ph.D. under the direction of Dr. Jeffrey D. Hart in May, 2005. Juhee Song married Jongil Lim. Her permanent address is 808-9 Woodang APT # 302, Mok-Dong Yangchun-Gu, Seoul, Korea.